

THERE IS NO LOGIC THAT CAN BE SUPERIMPOSED ON THE CITY; PEOPLE MAKE IT, AND IT IS TO THEM, NOT BUILDINGS, THAT WE MUST FIT OUR PLANS.

JANE JACOBS, *THE DEATH AND LIFE OF GREAT AMERICAN CITIES*

INFORMATION IS THE OIL OF THE 21ST CENTURY, AND ANALYTICS IS THE COMBUSTION ENGINE

PETER SONDERGAARD, *SVP, GARTNER*

ERRORS USING INADEQUATE DATA ARE MUCH LESS THAN THOSE USING NO DATA AT ALL

CHARLES BABBAGE *THE EDUCATION INDUSTRY*



BALAMURUGAN SOUNDARARAJ

# ESTIMATING FOOTFALL FROM PASSIVE WIFI SIGNALS

CASE STUDY WITH SMART STREET SENSOR PROJECT

DOCTOR OF PHILOSOPHY  
UNIVERSITY COLLEGE LONDON - UCL

DOCTOR OF PHILOSOPHY  
DEPARTMENT OF GEOGRAPHY, UCL

I, *Balamurugan Soundararaj* confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

.....  
*Submitted on, May 2019*

# *Contents*

1	<i>Introduction</i>	17
2	<i>Review of Literature</i>	19
3	<i>Collecting Wi-Fi Data</i>	37
4	<i>Processing the Data into Footfall</i>	51
5	<i>Visualisations and Applications</i>	81
6	<i>Discussion and Conclusions</i>	83
7	<i>Appendix</i>	85
8	<i>Research Article</i>	97
	<i>Bibliography</i>	113



# List of Figures

- 2.1 Growth of research in the areas of understanding distribution and dynamics of human activity since 1980. 20
- 2.2 Tree-map showing the volume of research conducted under each major themes and their sub-themes. 21
- 2.3 Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data 26
- 2.4 Distribution of research across various techniques and technologies 27
- 2.5 Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data 28
  
- 3.1 Structure of a probe request frame. *Source: IEEE 802.11 specification* 39
- 3.2 Number of probe requests collected every minute on 15 October 2017 41
- 3.3 Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data 43
- 3.4 Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data 43
- 3.5 Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data 45
- 3.6 The decay of signal strength (RSSI) with respect to distance. 47
- 3.7 Increase in the share of randomised MAC addresses compared to non-randomised original ones over the years. 48
- 3.8 Smartphone penetration by age group in United Kingdom (2012-18) 48
  
- 4.1 Comparison of volumes of datasets across various disciplines. 57
- 4.2 Missing data from five locations at Tottenham Court Road, London on 15 January 2018 demonstrating the veracity of the data. 58
- 4.3 Number of probe requests collected for every five minute interval at Tottenham Court Road, London on the year 2018 showing the visual complexity of data in the time dimension. 59
- 4.4 Big data characteristics of the Wi-Fi probe request datasets in their corresponding dimensions 60
- 4.5 Characteristics of types of Wi-Fi data collection tools at each end of the spectrum compared to an ideal candidate 61
- 4.6 Exponential increase in the processing time when using traditional methods. 63
- 4.7 The increase in processing time with the Unix pipeline is linear thus improves the scalability compared to R based processing 65
- 4.8 The scalability of the processing pipeline could be further improved with parallelising it. 65

4.9 Outline of the ‘Medium data toolkit’ devised to collect, process, visualise  
and manage the Wi-Fi probe requests data 67

# *List of Tables*

2.1	Evaluation of different technologies or approaches that can be used for data collection.	34
3.1	Significant information included in a probe request	39
3.2	Locations of installed sensors	43
3.3	Summary of the collected data-sets.	46
4.1	Comparison of volume or size of the datasets of Wi-Fi probe requests.	56
4.2	Comparison of velocity or speed of the datasets of Wi-Fi probe requests.	57
4.3	Examples of different types of Wi-Fi based data collection solutions.	61
4.4	Various data storage approaches and their characteristics.	63
4.5	Various types of big data processing tools and corresponding examples.	64
4.6	Tasks in the processing pipeline, corresponding R libraries and equivalent Unix tools	65
4.7	Comparison of clustering algorithms with a sample of 40000 probe requests	74
4.8	Locations where sensors were installs, volume and speed of probe requests collected by the sendor and total pedestrians manually counted. The data occupies around 1.8 GB on disk when encoded in text format.	76
4.9	Results of footfall estimation at each location as Mean Absolute Percentage Error (MAPE) after each step of the filtering process	78



# *Outputs From The Research*

## *Publications*

**Book Chapter** - Murcio, R., Soundararaj, B., & Lugomer, K. (2018). Movements in Cities: Footfall and its Spatio-Temporal Distribution. In Longley P., Cheshire J., & Singleton A. (Authors), *Consumer Data Research* (pp. 84-95). London: UCL Press.

**Journal Paper** - Soundararaj, B., Cheshire, J., & Longley, P. (2019) Estimating real-time high-street footfall from Wi-Fi probe requests, *International Journal of Geographical Information Science*.

**Conference Paper** - Soundararaj, B., Cheshire, J., & Longley, P. (2019). Medium Data Toolkit - A Case study on Smart Street Sensor Project. *In Proceedings of GIS Research UK*, New Castle, United Kingdom.

**Conference Paper** - Lugomer, K., Soundararaj, B., Murcio, R., Cheshire, J., & Longley, P. (2017). Understanding sources of measurement error in the Wi-Fi sensor data in the Smart City. *In Proceedings of GIS Research UK*, Manchester, United Kingdom.

**Conference Paper** - Soundararaj, B., & Zhu, D. (2019). Estimating pedestrian flow from footfall counts using Geo-propagation. *In Proceedings of Conference on Complex Systems*, Singapore. \*\*under consideration

**Conference Paper** - Murcio, R., Soundararaj, B., & Lugomer, K. (2018). Trends in urban flows: An information theory approach. *In Proceedings of Conference on Complex Systems*, Cancun, Mexico.

**Conference Paper** - Soundararaj, B., Murcio, R., & Lugomer, K. (2016). Smart Street Sensor Footfall Project. *In Proceedings of Conference on Complex Systems*, Amsterdam, Netherlands.

**Conference Paper** - Soundararaj, B., & Lugomer, K. (2016). Smart Street Sensor Footfall Project. *In Proceedings of Oxford Retail Futures Conference*, Oxford, United Kingdom.

## *Engagement Activities*

**Event Talk** - Estimating real-time high street footfall from Wi-Fi probe requests. (2018). *Geo+data*, London, United Kingdom.

**Event Talk** - Estimating real-time high street footfall from Wi-Fi probe requests. (2018). *Data Natives*, London, United Kingdom.

**Conference Workshop** - Estimating real-time high street footfall from Wi-Fi probe requests. (2018). *Smart Urban Policy Futures Workshop*, London, United Kingdom.

**Exhibition** - Real-time footfall in Bloomsbury. (2017). *Big Data Here - Big Data Week 2016*, London, United Kingdom.

### *Products*

**Research Dataset** - Smart Street Sensor footfall data, *Consumer Data Research Center*. URL: <https://data.cdrc.ac.uk>.

**Software** - Soundararaj, B. (2018). Clicker - an Android app for manually counting pedestrian footfalls with precision and accuracy. *Google Play Store*. URL: <https://play.google.com/store/apps?id=com.bala.manualcount>.

### *Awards*

**Bursary** - Yusuf Ali travel bursary (2017) *University College London*.

*Dedicated to my parents,  
S. Kalavathy and K. Soundararaj.*



# *Prologue*

## *Abstract*

- data is everywhere but no information - need to understand population in detail - this research connects this two quests - Wi-Fi probes in built environment = detailed footfall information - we conduct experiments collect two sets of data. - we process the data with different methods to get footfall - we establish the application of the data with series of examples

## *Impact Statement*

- we live in era of explosion of data. - everyone is looking for ways to use data - this research looks in to one such data set - deals with collection, processing and application. - we develop a open-source toolkit for doing all the above - the research resulted in publications - Outputs with C.D.R.C and subsequent use of the data - methods communicated to data partner for industry application

## *Acknowledgements*

- Supervisors - Family and Friends - Colleagues - Funding agencies - Open-Source projects (esp. GNU Linux and R)



# 1

## *Introduction*

We talk about the theory of cities and built environment. We start from how these have been perceived as function of the form and gradually changed to people, activity, economy and information. Built environment is manifestation of information exchange that happens in them. We talk about the change in theory regarding this. We talk about how this information exchange has been becoming more and more open and the opportunity it provides us planners, geographers and researchers to understand these things better.

Along with the information age there is an explosion of open data. The data collection has changed from structured high effort activity to low effort scraping activity. The data generated by scraping is unprecedented and staggering. Most of the Big-data research has gone into this in the past decade. Disadvantages of structured data which the unstructured data fills. This is changing how we view, understand and experience the world. Some of these datasets fall into this unique medium size category as well which are neither big data nor trivial. There is a need for methods and tools to collect, convert and use these data.

Talk about the ubiquity of the mobile technology. Everyone has a device which connects them with world wirelessly. Major ones cellphone and Wi-Fi. Wi-Fi is uniquely placed in between Cellphone and Bluetooth. The design of Wi-Fi gives us amazing opportunity. This has been done before for the past decade by loads. The privacy advocacy has become a new thing. The change is from both ends. Collectors are regulated, cellphones are getting sneakier. Need for method to collect data and analyse it without compromising on privacy. The conversion of this unstructured data into something tangible and measurable is not a trivial problem. There are loads of such data and measurements. Examples - banking vs economic activity, oyster card data vs movement.

The potential use of such information is immense. Give examples of smart city paradigm and connected city where real time census is possible. We can not only take snapshot of the state of the city, we can record and understand the built environment as living, breathing organism. The insights we get by combining this information with other similar info is more than sum of their parts. It can revolutionise understanding , planning, policy etc, urban management and finally industry such as retail, transportation etc. city mapper, sharing economy etc.



## 2

### *Review of Literature*

Understanding the scale, nature and dynamics of distribution of population in space and time has been a central premise of academic research in various fields of study such as human geography, sociology, urban planning and architecture. This granular knowledge of where people are and how they move is critical in practical decision making in various industries such as real estate for valuing places, retail for business planning and emergencies for risk management and evacuation. The first challenge in any of these research is collecting precise and accurate data. This started as research into methods estimating and interpolating highly granular data from existing regional level aggregate datasets and as the technology improved through the later half of twentieth century, the research methodologies adopted the new technologies to utilise these more granular sources of data. Though new technologies provide immense opportunity in collecting large amounts of data which were previously impossible they also introduce their share of uncertainties. Hence it becomes imperative that we understand the evolution of these techniques and methodologies along with the research that used them to build our rationale behind any further research.

Moreover with the proliferation of mobile devices and wireless internet connectivity, even day to day activities are being digitised leading to the creation of large amount of easily accessible data which are generated passively in an unstructured manner. The users' acceptance to the collection and analysis of such data has also been improving until recently <sup>1</sup>. There has also been rising concerns regarding user privacy along with the development of more accurate methods to track them. In this context, the critical task in all these research is to solve the problem of balancing these two by choosing the right technologies and devising the appropriate methods.

In this chapter we conduct a systematic survey of literature in this broad area of 'distribution and dynamics of human activity'. The aim of this survey is to evaluate where we are at the research and understand how we got here. First we conduct a comprehensive survey of over 300 publications which discuss this area of research and we then look at the major themes and trends in the last 70 years. We discuss these themes in detail looking at the aims and achievement of these research while highlighting the opportunities and gaps they leave. We also look at the timeline of their publication and discuss the evolution of the research along with the changes in the technology landscape. We then classify the studies by looking at the technologies employed by them and note the trends in the adoption and phasing out of relevant technologies. We also

<sup>1</sup> Alfred Kobsa. User acceptance of footfall analytics with aggregated and anonymized mobile phone data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8647 LNCS, pages 168–179, 2014. ISBN 9783319097695. DOI: 10.1007/978-3-319-09770-1\_15

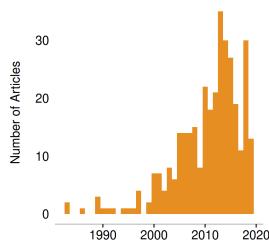


Figure 2.1: Growth of research in the areas of understanding distribution and dynamics of human activity since 1980.

*Measured in the number of papers published*

<sup>2</sup> Donald L Foley. Urban daytime population: a field for demographic-ecological analysis. *Social Forces*, pages 323–330, 1954. ISSN 0037-7732; and Robert C Schmitt. Estimating daytime populations. *Journal of the American Institute of Planners*, 22(2): 83–85, 1956. ISSN 0002-8991

try and understand the advantages and disadvantages of these techniques and develop a theoretical framework to understand when and how to use them effectively answer research questions. Finally we summarise the literature survey focussing on the major gaps still left in our understanding and the new, emerging areas where more research is necessary. We also critically evaluate these areas of research in terms of priority and feasibility to develop our research questions and devise a plan for our research.

We begin our literature review with a systematic search for academic publications studying the distribution and dynamics of human activity. We start this search from a set of reviews pertaining to the topic such as Ratti et al. [2006], Jiang et al. [2013], Steenbruggen et al. [2013], Arribas-Bel [2014] and Li et al. [2016]. From these reviews we further expand by navigating the citation networks and finding research which are relevant to our search. Though this does not give us a perfectly comprehensive set of research literature, it does provide us with a representative sample of all the different directions of the research conducted in the area. Through this process, we identified around 325 relevant research publications which deal with the collection, measurement, analysis, visualisation and discussion of population at a granular level. The research started around 1950s looking at estimating day-time urban population at a granular level using existing broader data employing various estimation methods <sup>2</sup>. Though this served as a starting point, the pursuit of such granular data and their applications in corresponding fields didn't pick up until the start of the 21st century fuelled by the digital revolution that followed growth of internet. Figure 2.4 shows the yearly volume of research published since 1980. We can observe that though there were some research conducted through 80s and 90s the real push forward came around beginning of the millennium when mobile phones adoption skyrocketed. In addition to the early 2000s we also see a substantial increase in interest in the beginning of the next decade fuelled by the smartphone revolution which completely changed the research avenues in-terms of volume and types of data available and methodologies available to tackle them. The area of research is multi-disciplinary encompassing academic interest and commercial applications in Geography, Urban Studies, Urban Planning and Management, Emergency planning and Management, Economics, Computer Science and Engineering etc.

## 2.1 Research Themes

In this section we look at the major themes and questions tackled by this knowledge base. We start by classifying the research into the major and minor themes explored in them as shown in Figure 2.2. The tree-map shows the volume of research in corresponding themes measured in terms of number of publications. We can observe that the research is conducted in five major areas - population studies focussing on the creating and utilising data on distribution and nature of human activity, mobility and interaction focussing on the changes in these distributions, understanding the nature and function of space from these distribution and change, methods and techniques which can be used to conduct the research and finally issues and solutions related to the privacy of the users while conducting these research. We can also observe

that most of the research apart from developing methods were conducted in the domain of human mobility and social interaction closely followed by the population distribution. In the following sections we discuss these in detail along with their sub themes with the following framework,

1. What are the major lines of questioning?
2. What has been done previously?
3. Where are the opportunities for further research?

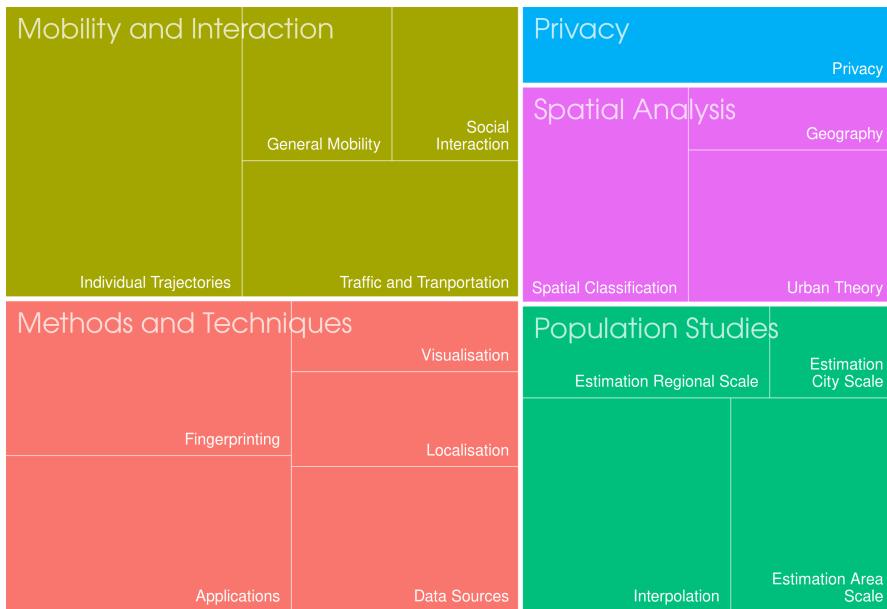


Figure 2.2: Tree-map showing the volume of research conducted under each major themes and their sub-themes.

### 2.1.1 Population Studies

Though Foley [1954] and Schmitt [1956] started this line of research in 1950's with the discussion on estimating daytime population using broader datasets it was not until the 80s significant volume of research kicked off in this area of study. From 80s until mid 2000's numerous studies were conducted on measuring and studying the population at a granular level both spatially and temporally. The focus of the research around this time was primarily on interpolation from the larger datasets created using censuses, regional or national level sample surveys and other centrally collected sources of data. There have been numerous fairly successful attempts with methodologies where a broad dataset such as regional level population summaries and modelling or interpolating more granular data from them by augmenting with other sources of data such as street networks [Reibel and Bufalino, 2005], remote sensing [Sutton, 1997, Yuan et al., 1997, Chen, 2002] etc. Dobson et al. [2000, 2003], Bhaduri et al. [2002, 2007] and [Mennis, 2003, Mennis and Hultgren, 2006] are examples of such research methodology. These studies were almost done on a city scale or above with mostly modelling or interpolation methods since the data sources were few and were centrally collected.

Around 2005, there was a sharp shift in research where the interpolation methods were replaced by highly available granular data collected over cellular network. Studies were conducted on estimating population densities, presence of tourists, general activity patterns using data from cellular networks. Most of these research were conducted at a far larger geographic scale looking at things at an area level [Pulselli et al., 2008, Girardin et al., 2009, Phithakkitnukoon et al., 2010, Yuan and Raubal, 2016]. There were efforts in using device level sensors such as global positioning system(GPS), Wi-Fi and Bluetooth to detect population distribution and socio-geographic routines [Calabrese et al., 2010, Rose and Welsh, 2010, Farrahi and Gatica-Perez, 2010]. There have been studies on looking at people distribution as granular as queue lengths as discussed by [Wang et al., 2013] to city level dynamic population mapping where the limitations of traditional datasets generated through censuses and surveys<sup>3</sup>.

<sup>3</sup> Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014. ISSN 0027-8424. doi: 10.1073/pnas.1408439111

Around the 2015, along with the data collected directly from the mobile devices, the data that are generated by the users activity on these devices are became more important. Social media data such as twitter [Lansley and Longley, 2016b] and other consumer data such as loyalty cards [Lloyd and Cheshire, 2018], smart cards [Ordonez and Erath, 2012] etc. have also become a significant sources of data for such research. Recently, with increased concerns and legislation on privacy, there have been studies which go back to the effort of interpolating granular data from broader datasets but using more data and processor intensive technologies such as agent based modelling, deep learning, small area estimation [Crols and Malleson, 2019, Shibata and Yamamoto, 2019, Rao and Molina, 2015] etc.. Though there have been a lot of work done in most of the directions in this research area, the clear gap arises due to the absence of a continuous, granular and sufficiently longitudinal data-sets to complement the methodologies that have been developed.

### 2.1.2 Human Mobility and Interaction

<sup>4</sup> M Castells. Grassrooting the space of flows. i wheeler, aoyama and warf [eds.] cities in the telecommunications age, 2000

Study of movement of people is one of the major areas of research which have significantly benefited from the decentralised collection of data at a granular level<sup>4</sup>. In addition to being useful in their own right, these data were in turn used to augment traditional models of travel behaviour, traffic and transport to provide a better understanding of human movement over time and space [Janssens, 2013]. The major themes of research within this area are, Movement of people in space and time with emphasis on understanding the built environment, social interaction between these people with a sociology perspective and traffic and transportation studies with a infrastructure perspective. There is significant volume of research which dealt with recording and analysing the trajectories of the users to understand their movement patterns enabled by the unprecedented availability of detailed data from mobile devices and this is discussed in detail along with the discussion of the technologies used in Section 2.3.

### 2.1.3 Methodology and Techniques

Research in this are focused around 5 major topics,

1. Localisation - Research into using the mobile devices and data generated from them as a cheaper alternative to Global Positioning Systems and remote sensing.
2. Data Sources - Identifying and formalising new data-sources as the technology develops
3. Application - Applying these identified data sources to answer questions and solve problems in different disciplines.
4. Visualisation - simplifying, visualising and interpreting these high volume of unstructured, noisy datasets.
5. Device fingerprinting - overcoming the difficulties posed by the anonymisation process and extract useful information.

Localisation of mobile devices without the use of expensive additional infrastructure such as GPS is one of the earliest ideas pursued in this aspect [Bulusu et al., 2000, He et al., 2003, Moore et al., 2004, LaMarca et al., 2005]. This research, when reversed, could also lead to the tracking of these devices in space without the aforementioned infrastructure thus providing a inexpensive, easy way to collect mobility data. The sensors which are already present in the phones such as Bluetooth [Bandara et al., 2004], Wi-Fi [Zarimpas et al., 2006], cellular radio [Dil and Havinga, 2011, Ahas and Mark, 2005] etc. have been considered to be used for localisation of the devices. This has been particularly important in the field of indoor localisation where GPS doesn't usually work<sup>5</sup>. When seen from the other perspective the same technologies and methods can enable us to collect presence and movement data on people indoors [Roy and Chowdhury, 2018a,b, Jia et al., 2019, Nikitin et al., 2019, Kulshrestha et al., 2019, Deng et al., 2018].

The identification of data sources started with looking at the 'real time' city examining the digital landscape created by the citizens their electronic devices [Townsend, 2000]. This was furthered by the notion of 'instrumenting' the city and developing methods and techniques under the umbrella of smart cities and internet of things [O'Neill et al., 2006, Sruthi, 2019]. Since there have been research looking at the wireless data collected from positioning technologies [Bensky, 2007] and cellular network [Kiukkonen et al., 2010, Steenbruggen et al., 2015] and even crowdsourcing as method of collection [Shin et al., 2013] leading towards a framework for computational urban planning [Kontokosta, 2015]. With the effort to formalise them as valid sources of data, there have also been research looking at the biases in them such as mobile phone ownership [Wesolowski et al., 2013, Kobus et al., 2013].

Identifying and fingerprinting unique devices and users from noisy, unstructured data is another area of active research under methodologies and techniques<sup>6</sup>. The majority of the work has been done as an extension of localisation where the GPS-less positioning leading to finger printing people and their movement out of the data [Pang et al., 2007a, Pappalardo et al., 2015]. Additionally there are work looking at the tracks collected from Wi-Fi or mobile data and extract unique users out of them [Girardin et al., 2008, Eagle and Pentland, 2009, Jiang et al., 2012]. It is also demonstrated that it is possible to wireless technologies can be used to detect even device free entities [Elgohary, 2013]. These localisation and clustering techniques can also be used for socio-geographical analysis and to understand the patterns of

<sup>5</sup> Nobuo Kawaguchi. Wifi location information system for both indoors and outdoors. In *International Work-Conference on Artificial Neural Networks*, pages 638–645. Springer, 2009

<sup>6</sup> Bin Jiang and Xiaobai Yao. Location-based services and gis in perspective. *Computers, Environment and Urban Systems*, 30(6):712–725, 2006. ISSN 0198-9715; and Lin Liao. *Location-based activity recognition*. PhD thesis, University of Washington, 2006

activity of people [Licoppe et al., 2008]. There have been quite a few security research on the robustness of the anonymisation techniques while revealing methodologies to overcome limitations imposed by them [Mathieu Cunche, 2016, Chothia and Smirnov, 2010, Krumm, 2007]. Cheng and Wang [2016] was one of the first to look into devising a method to do this in a non-intrusive way which are further extended by Di Luzio et al. [2016], Adamsky et al. [2018] and Dai et al. [2019]. This is currently an active field of research and there is immense opportunity for further research.

Visualising the temporal dynamics of data collected on human activities through decentralised processes poses significant challenges when approached with traditional cartographic concepts <sup>7</sup>. Digital media especially animation has been explored as an option to solve for the temporal dimension [Morrison et al., 2000, Lobben, 2003] but is bound by the cognitive limits of the viewer [Harrower, 2007]. There have been approaches proposed around animations of generated surfaces [Kobayashi et al., 2011] and network-based visualizations [Ferrara et al., 2014] leaving gaps in research for new methods in dynamic geographic visualisation [Fabrikant, 2005] and visualising path and flow of phenomena [Thomas, 2005], particularly of the mobility data collected from cellphones [Sbodio et al., 2014]. This provides us with a promising opportunity for research in methods for visualising high frequency, hyper-local pedestrian data within the limits of cognition of the viewer.

#### 2.1.4 Spatial Analysis - Theory and Modelling

Traditional and modern geography was dominated by the study of centrally collected data acquired through extensive field surveys and remote sensing. In the last two decades, a significant paradigm change has been introduced by the availability of unprecedented amount of data generated by unconventional sources such as mobile phones, social media posts etc. This move to the post-modern geography has been accompanied by a change in our understanding of the built environment and human geography from a static point of view to a more dynamic definition <sup>8</sup>. This definition is based on the bottom-up mechanisms which make human activity such as information exchange and economy to manifest in the physical built environments as argued by [Batty, 1990, 1997, Batty et al., 2012] and [Batty, 2013a,b].

This transition into the digital age [Graham and Healey, 1999, Tranos and Nijkamp, 2012, Tranos, 2013] has changed the politics of space and time [Massey, 1992] and been more pronounced in the study of urban built environment where technology has redefined the concepts of place and space [Graham and Marvin, 2001, 2002, Sassen, 2001]. With the ability to collect and analyse of data on large complex systems in real-time [Graham, 1997], we are exploring the possibilities of understanding their structure and organisation using concepts of complexity theory [Bettencourt, 2013, Portugali et al., 2012] with more emphasis on their temporal patterns such as the argument towards finding the pulse of the city [Batty, 2010]. With the population getting more and more connected [Castells, 2010], the nature of space/place is being dynamically defined by the population themselves [Giuliano and Small, 1991] and vice versa [Zandvliet and Dijst, 2006]. This flood of hard data <sup>9</sup> was accompanied not only by optimism in its potential [Thomas, 2001] but also

<sup>7</sup> Alan M MacEachren and Menno-Jan Kraak. Research challenges in geovisualization. *Cartography and Geographic Information Science*, 28(1):3–12, 2001. ISSN 1523-0406; and Elaine J Hallisey. Cartographic visualization: an assessment and epistemological review\*. *The Professional Geographer*, 57(3):350–364, 2005. ISSN 0033-0124

<sup>8</sup> Edward Soja. Postmodern geographies, 1989. URL <http://books.google.com/books?id=sNcRAQAAIAAJ>

doi: 10.1038/453698a

by the questions raised on the challenges in handling the diverse, large scale, non standardised data it produces and the usefulness or representativeness of the resulting analysis [Miller, 2010, Arribas-Bel and Sanz-Gracia, 2014].

However, availability of such data has impressive uses in urban studies <sup>10</sup> especially with advancement of new technologies [Steenbruggen et al., 2013] and possibility of distributed, crowdsourced data collection [Lokanathan and Gunaratne, 2015].

### 2.1.5 Privacy

The ubiquity of personal devices and digitisation of day to day activities through these mobile devices [McMeel, 2018] has provided many opportunities for researchers and industry for collecting, analysing and deriving inputs from them. However at the same this also increased the risk of infringement on privacy of the users whose data is being collected <sup>11</sup>. There is immense value in uniquely identifying and profiling information on people for specialised purposes such as security [Cutter et al., 2006] and law enforcement [Dobson and Fisher, 2003] but also has extreme risks associated when not handled with care [VanWey et al., 2005].

Strictly protecting personal information while ensuring the information is usable for research by maintaining the uniqueness in the data is the major concern which was addressed by devising frameworks for secure practices in confidentially collecting and using the location data [Duckham and Kulik, 2006, Tang et al., 2006, Lane et al., 2014]. Some efforts sought to accomplish this task through cryptographic hashing algorithms (Pang, 2007) while others aimed to thwart identification and tracking at the device level by techniques such as MAC randomisation [Gruteser and Grunwald, 2005, Greenstein et al., 2008]. Finally though getting consent of users for the collection and use of such information from their mobile devices is challenging, there is a significantly improved acceptance when the process offers value in return such as discounts and monetary benefits [Kobsa, 2014].

There is opportunity in this area for research in applying the cryptographic solutions along with the privacy preserving frameworks to arrive at methods which can extract useful information out of large personal data while obscuring or anonymising them.

## 2.2 Research Trends

Figure 2.3 shows the volume of research done in this topic since 1980 categorised based on their major themes discussed earlier. We can observe that there are distinct trends in the research over time, which evolved around the development of technology in the last two decades. Until 90s the research was mostly centered around population studies on estimating and interpolating granular spatial and temporal information from larger and cross sectional datasets such as census and sample surveys. The period between 2000-2010 there was interest in potential of the new data generated by the digital revolution. We can categorise this as the ‘mobile era’ where carrying mobile devices become mainstream. This explosion of research coincided with mobile phones becoming more popular and ubiquitous with population in urban

<sup>10</sup> Luis M A Luís M.A. Bettencourt. The uses of big data in cities. *Big Data*, 2(1):12–22, 2014. ISSN 2167-6461. DOI: 10.1089/big.2013.0042. URL <http://online.liebertpub.com/doi/abs/10.1089/big.2013.0042>

<sup>11</sup> T Scott Saponas, Jonathan Lester, Carl Hartung, Sameer Agarwal, and Tadayoshi Kohno. Devices that tell on you: Privacy trends in consumer ubiquitous computing. In *Usenix Security*, volume 3, page 3, 2007; and John Krumm. A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009

areas and was around development of methods and techniques to utilise the data generated from them. There were also extensive studies in using the datasets to understand human mobility along with a rising concern in the privacy of the users who's data which are being used for these studies.

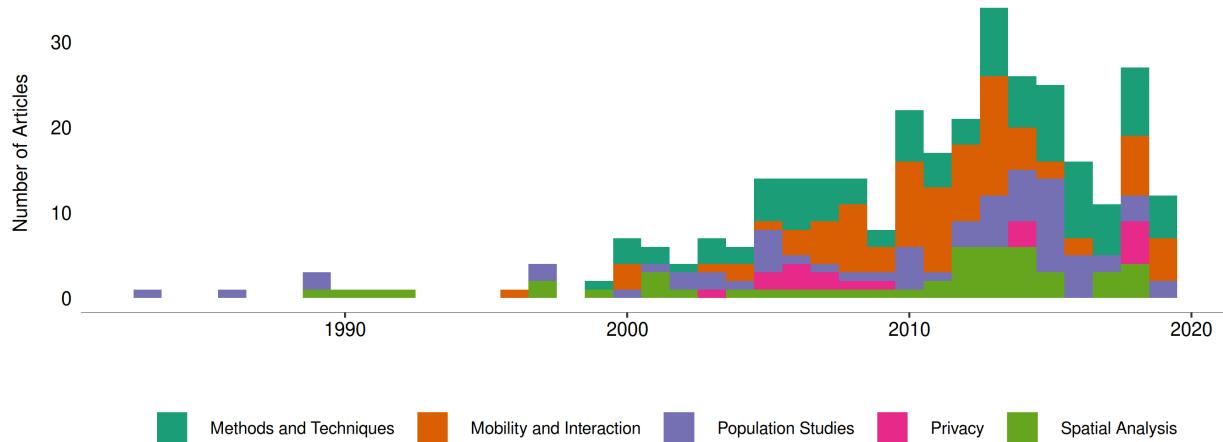


Figure 2.3: Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data

The release of iPhone in 2008 and the increase in the share of 'smartphones' in the next 10 years sparked the 'smartphone' era. The change made sure that all the mobile devices gaining numerous capabilities such as internet connectivity over Wi-Fi and mobile network, location awareness with global positioning system, movement recognition with accelerometers and connectivity other 'wearable' devices through Bluetooth. This also lead to the digitisation of lifestyle where every aspect of the life being done through these devices over internet while generating huge amount of data on these activities. This sparked the large volume of research on the form and function of space by studying this data and on the dynamics of human population in space and time in the next 5 years. These research were particularly centered around tracking the trajectory of people using the mobile devices they carry with them as the smartphones made it easier to collect the necessary data directly from them rather than depending on a centrally collected datasets from mobile carriers. With the theoretical limit to predictability in human mobility quantified by [Song et al. \[2010b\]](#), the focus on urban mobility has been declining in the past few years which has led to a renewed interest in population studies at a local-local level in real-time. In addition to using the data from the mobile devices, these studies have also been exploring the use of large assemblages of consumer data that are being generated in this connected mobile environment and linking them together to create a fuller picture <sup>12</sup>

<sup>12</sup> Paul Longley, James Cheshire, and Alex Singleton. *Consumer Data Research*. UCL Press, 2018

Finally, with the increase in use of personal data, there has also been an increase in research regarding the privacy of the users. Along with this, the mobile devices and subsequently the data generated by them are more and more anonymised so that the users cannot be tracked or identified at a personal level. This has given rise to the new trend in research to devise methods to overcome this anonymisation and at the same time research which considers these methods as vulnerabilities and find solutions to make the anonymisation process more robust. There is clear need for methods which

anonymise the data sufficiently to protect the identity of the users and at the same time enable us to conduct research in measuring studying population distribution and movement at a granular level.

### 2.3 Techniques and technology

When we look at the literature from the technology perspective, we observe that over the years, the research continuously picks up and applies recent technological developments in the pursuit of understanding the distribution of human activity and population. Figure 3.5 shows the distribution of the research in terms of the main technique/ technology used over the past 50 years. We observe that the earliest attempts started from the exploration of using interpolation and modelling techniques on a broader dataset. As the need for more granular datasets increased there were attempts to devise and utilize bespoke solutions to generate them. When mobile devices became mainstream, the focus shifted to utilize the relevant components of the mobile infrastructure. A significant number of studies were done in utilising data collected from the mobile network, sensors in the mobile devices, especially GPS and Wi-Fi, in addition to the social media content generated from these devices. A detailed account of these studies is given below,

#### 2.3.1 Interpolation and Modelling

Attempts in using the existing data collected through traditional methods such as census and large scale sample surveys to create spatially and temporally granular and detailed estimates were carried out by applying various interpolation methods such as pycnophylactic, dasymetric interpolation [Tobler, 1979, Mennis, 2003, Mennis and Hultgren, 2006, Hawley and Moellering, 2005, Tapp, 2010, Wismans et al., 2017] along with spatial [Lam, 1983, Martin, 1989, Martin et al., 2015] and temporal interpolation techniques [Glickman, 1986]. These methods along with supplementary data such as remote sensing imagery [Sutton et al., 2001, Chen, 2002] and street networks [Reibel and Bufalino, 2005] were shown to be useful in producing detailed granular population maps at various scales with varying degree of success [Dobson et al., 2000, Bhaduri et al., 2002, Dobson and Fisher, 2003, Bhaduri et al., 2005, 2007]. These approaches have been employed in various applications such as econometric studies [McDonald, 1989], studies on public health [Hay et al., 2005], emergency management [Kwan and Lee, 2005] and flood risk estimations [Smith et al., 2016].

In addition to these interpolation techniques classic modelling techniques can also be used to estimate daytime populations and demographic structure at hyper-local scales [Jochem et al., 2013, Jia et al., 2014], urban scales [Alahmadi et al., 2013, Abowd et al., 2004] and regional scales [Foley, 1954, Schmitt, 1956, Singleton and Longley, 2015, McCormack, 2017]. The granular data created with such modelling techniques are shown to be useful in urban planning and management [Parrott and Stutz, 1999], emergency management [Alexander, 2002, Cutter et al., 2006] and in modelling traffic and transportation [Lefebvre et al., 2013]. These interpolation and modelling techniques along with granular data produced are also used in classifying spatial areas

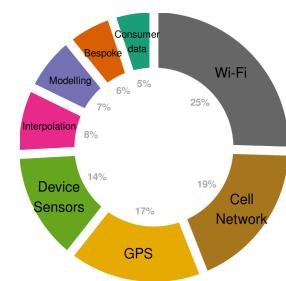


Figure 2.4: Distribution of research across various techniques and technologies

*Measured in the number of papers published*

and hence understanding the structure of cities in general [McMillen, 2001, 2004, Lee, 2007, Arribas-Bel, 2014]. Though being useful, these techniques are still shown to have limitations and uncertainties [Nagle et al., 2014], which mostly arise from the nature of the input data employed. This leads us to the need for more detailed and frequent collection of data.

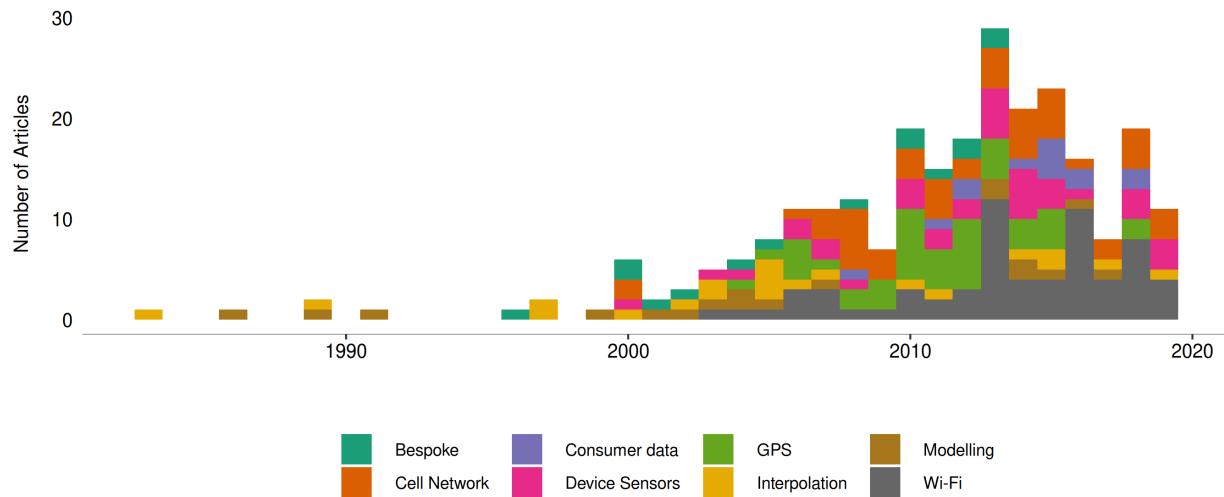


Figure 2.5: Outline of the ‘Medium data toolkit’ devised to collect, process, visualise and manage the Wi-Fi probe requests data

### 2.3.2 Bespoke technologies

Following this need, there has been efforts to use bespoke or specialised technologies such as cameras [Cai and Aggarwal, 1996, Heikkilä and Silvén, 2004, Kröckel and Bodendorf, 2012], Lasers [Zhao and Shibasaki, 2005, Arras et al., 2008] and radio frequency receivers [Bahl and Padmanabhan, 2000, Yang et al., 2013, Chothia and Smirnov, 2010, Bulusu et al., 2000, Dil and Havinga, 2011] to measure human activity. But the major problem with such solutions is the cost and effort involved in designing and implementing them at urban and regional scales comprehensively. Moreover, being specialised and centralised they tend to be challenging to maintain and update as the technological landscape change. This gives us the need to identify and use techniques which are more general in nature and can be used for longer periods of time which are cheap to install to achieve a more comprehensive coverage.

### 2.3.3 Cellular Network

The rise of mobile phones as ubiquitous personal devices for the broader population has provided us with a viable alternative for collecting data with greater granularity at large scales. Mobile infrastructure consists of both the ‘network part’, built and managed by the service providers, and the ‘user part’, which is the phones owned by the users’ themselves. The network part, in addition to providing connectivity to the users, also collects information on these devices actively such as communication between the users and passively such as when the phones themselves move from tower to tower. The mobile devices themselves have a variety of sensors such as accelerometer to identify

movement, compass to identify orientation, GPS receiver to deduct geographic position, etc. They also have various communication capabilities such as cellular, Wi-Fi, Bluetooth and Near field communications (NFC) etc. Both of these sensors and communication capabilities can be used as sources of data themselves. With the growth of mobile devices and the infrastructure surrounding it, there has been significant effort in utilising data generated by every component of this complex infrastructure.

The first set of research started to use the cellular network data for urban studies [Jiang et al., 2013, Steenbruggen et al., 2015, Lokanathan and Gunnarathne, 2015, Calabrese et al., 2015, Reades et al., 2007]. Even though this approach has been acknowledged to have inherent biases such as ownership bias across particular demographic groups [Wesolowski et al., 2013] the relative advantages such as coverage made them excellent sources of data. Visual exploration of use of such data using interactive interfaces to evaluate quality of service and scenario testing has been tested for the optimisation of public transport [Sbodio et al., 2014]. Such network data with the active and passive information collected from them can be used to create trajectories of people<sup>13</sup>, detect their daily routine [Sevtsuk and Ratti, 2010] and classify those routes in terms of function [Becker et al., 2011a]. It was also demonstrated to be useful in understanding overall mobility and flow of people and information [Candia et al., 2008, Krings et al., 2009, Simini et al., 2012, Zhang et al., 2019]. These data can be used to identify asymmetry in flow of people spatially [Phithakkitnukoon and Ratti, 2011], estimate volume and pattern of road usage [Bolla et al., 2000, Wang et al., 2012] and by augmenting the topology to optimise operations [Puzis et al., 2013]. Such datasets have been extensively used in traffic and transportation research to derive origin-destination matrices [Caceres et al., 2007, Mellegard et al., 2011, Iqbal et al., 2014], travel time estimation [Janecek et al., 2012] and traffic status estimation [Demissie et al., 2013, Grauwin et al., 2015].

It has been shown that mobile network data can be used to uncover nature of the population such as tourists in specific areas [Girardin et al., 2008] and the interaction between the people in the study area [Campbell et al., 2008]. The structure [Onnela et al., 2007a,b], geography [Lambiotte et al., 2008] and dynamics [Hidalgo and Rodriguez-Sickert, 2008] of such networks have been studied and demonstrated to be useful in predicting their change [Wang et al., 2011, Vajakas et al., 2018] over time. This social networks and their spatio-temporal structure can also be used for classification of land use [Pei et al., 2014, Jia et al., 2018], assessment of spatial patterns [Reades et al., 2009, Steenbruggen et al., 2013] and understanding the broader spatial structure of cities [Louail et al., 2014, Arribas-Bel and Tranos, 2015] and regions [Arhipova et al., 2018]. The data collected from the cellular network when examined at granular levels such as inter-personal communication and economic activity can be used to create estimations of micro area-level population density [Pulselli et al., 2008, Ng et al., 2017] and also the characteristics [Girardin et al., 2009] and the nature of the activity [Phithakkitnukoon et al., 2010]. Aggregated human activity measured from such research in turn can be used to measure and model population dynamics and land use density and mix at broader level [Jacobs-Crisioni et al., 2014, Tranos and Nijkamp, 2015, Tranos et al., 2018]. The spatial patterns thus uncovered can then be applied to

<sup>13</sup> Johannes Schlaich, Thomas Otterstätter, and Markus Friedrich. Generating trajectories from mobile phone data. In *Proceedings of the 89th annual meeting compendium of papers, transportation research board of the national academies*, 2010

urban planning [Becker et al., 2011b] whilst the temporal patterns uncovered have immense utility for the disciplines such as epidemiology. For example, population influxes measured from changes in mobile network usage can be used to model spread of diseases [Buckee et al., 2015].

Though the mobile network provides much more granular and accurate data than interpolation techniques, it is not without its limitations [Yucel, 2017]. The distribution of network infrastructure usually follows the purposes of service coverage and follows commercial decisions. This introduces systematic biases in the data passively collected through them. Moreover, the data actively collected through them has bias based on the volume of usage of services by the customers which can vary widely spatially, temporally and also based on demography. In addition to this because of the coverage, the data collected from mobile service providers pose immense privacy risk when linked to other sources of consumer data. This makes collection of data directly from the devices using the sensors on the device much more robust in certain cases.

#### 2.3.4 Mobile Sensors

The most prominent sensors and capabilities present in mobile devices that can be used for distributed urban sensing are Cellular radio, Bluetooth, Wi-Fi, GPS, accelerometer and compass <sup>14</sup>. Since cellular radio is managed by the cellular network and covered in mobile network data, we explore the research done with other sensors. In contrast to planned actively collected data, data passively collected via a distributed network of general purpose devices tends to be larger and more temporally dynamic. For example, an organised survey conducted every month to understand interpersonal communications between people in a team of 50 will result in a 2500 records a month. The same task is done through collecting data on email communication sent by them will result in a same volume records in a day. The challenges and solutions on collecting and analysing such large-scale longitudinal data are discussed by [Laurila et al., 2012, Antonic et al., 2013]. The real time nature of such data also gives us the opportunity to monitor and understand the city in much smaller temporal scales [Townsend, 2000, O'Neill et al., 2006] and the representativeness of such datasets have also been explored [Shin et al., 2013, Kobus et al., 2013]. Data generated from communication networks can be used to understand the structure of urban systems which are becoming increasingly border-less <sup>15</sup>.

Similar to the network based data, it can help in understanding human mobility [Asgari et al., 2013, Amini et al., 2014, Zhang et al., 2014] through mining trajectory patterns [Giannotti et al., 2007] and socio geographic routines [Farrahi and Gatica-Perez, 2010]. It is also useful in various traffic and transportation applications for monitoring roads [Mohan et al., 2008] and estimating traffic [Cheng et al., 2006], uncovering regional characteristics [Chi et al., 2014] and extracting land use patterns [Shimosaka et al., 2014]. Apart from GPS and Wi-Fi, there have been efforts in exploring other possibilities such as Bluetooth for location [Bandara et al., 2004, Becker et al., 2019] and aggregate detected Bluetooth activity to monitor freeway status [Haghani et al., 2010]. There have also been successful implementations of frameworks to predict movement of people by combining Wi-Fi and Bluetooth [Vu et al., 2011]. But owing to shorter range and requirement of active engagement from

<sup>14</sup> Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. A survey of mobile phone sensing. *IEEE Communications magazine*, 48(9), 2010; and Enwei Zhu, Maham Khan, Philipp Kats, Shreya Santosh Bamne, and Stanislav Sobolevsky. Digital urban sensing: A multi-layered approach. *arXiv preprint arXiv:1809.01280*, 2018

<sup>15</sup> Luca Bertolini and Martin Dijst. Mobility environments and network cities. *Journal of urban design*, 8(1):27–43, 2003

the user where they have to actively start the device pairing process, Bluetooth is much less preferable for large-scale data collection than GPS or Wi-Fi. The research on GPS and Wi-Fi based studies are discussed in more detail below.

### 2.3.5 Global Positioning System

In addition to providing a user's location to applications such as maps and navigation, the GPS capability in mobile devices in tandem with Wi-Fi can also maintain a continuous list of locations visited by the device over long periods of time. It works mostly in the background and requires almost no active input from the user to operate. Though very convenient for collecting data, due to the privacy risks associated with it, GPS is often one of the resources in a device that requires explicit user permission to be accessed. The concepts and methodologies for collecting such data were set out by [Asakura and Hato, 2004] and there have been attempts to collect this rich data from volunteers at a large scale along with ancillary data [Kiukkonen et al., 2010] and provide a location based service application for the collection of data [Ratti et al., 2006, Jiang and Yao, 2006, Ahas and Mark, 2005].

The accuracy, convenience and being designed for navigation makes GPS one of the most used technologies for mobility studies<sup>16</sup>. It has been used to analyse and understand individual mobility patterns [Neuhaus, 2009, Shin et al., 2010], which have been shown to have a high order of regularity in spite of the complexity [Brockmann et al., 2006, Song et al., 2010a]. There have been efforts to use this regularity to predict the future location of people [Monreale et al., 2009, Calabrese et al., 2010]. The limitations of predictions have also been quantified [Song et al., 2010b]. There have been successful efforts in extracting behaviours and patterns from such trajectory data [Liu et al., 2010, Cho et al., 2011, Hoteit et al., 2013, Pappalardo et al., 2013] along to understand individual patterns from large assemblages [Giannotti et al., 2011, Calabrese et al., 2013] and vice versa [Wirz et al., 2012]. In traffic and transportation, GPS trajectory from mobile devices is used to estimate [Calabrese et al., 2011] and expand [Jing et al., 2011] OD matrices, detect the mode of travel [Gong et al., 2012, Rossi et al., 2015] and calibrate existing spatial interaction models [Yue et al., 2012]. Since the data is collected at the device level and depends on the activity of the individual, it can be de-anonymised to reveal the nature of the owner of the devices. The possibilities of detecting the activity of the individual from trajectory information is demonstrated by [Liao, 2006, Krumm, 2007]. Patterns [Jiang et al., 2012] and structures in routines [Eagle and Pentland, 2009] can be extracted from these trajectories and can be used for socio geographic analysis of the population [Licoppe et al., 2008, Chen et al., 2018]. It can also be utilised in classification of the population at a particular location at a given time [Pappalardo et al., 2015]. Being inherently spatial and activity driven, GPS trajectories have been shown to be useful to identify [Bao et al., 2012], characterise [Wan and Lin, 2013] and automatically label [Do and Gatica-Perez, 2014] significant places of interest. It can also be used for land use detection [Toole et al., 2012, Zhang et al., 2018], classification [Jiang et al., 2015] and the study of urban morphology [Kang et al., 2012]. These GPS trajectories have been shown to be useful in estimating population dynamics at local level and within short durations during social events [Calabrese et al.,

<sup>16</sup> Marta C. González, Cesar A César A. Hidalgo, Albert-László Barabási, Marta C Gonzalez, Cesar A César A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008. ISSN 0028-0836. DOI: 10.1038/nature06958. URL <http://www.nature.com/nature/journal/v453/n7196/full/nature06958.html>{}5Cnhttp://www.nature.com/nature/journal/v453/n7196/pdf/nature06958.pdf

2010, Kim and MacEachren, 2014, Deville et al., 2014]. When combined with other data sources can be useful to understand relationship between spatial areas [Long and Thill, 2015].

From the literature we see that GPS is one of the most precise and accurate user side methods of collecting location of mobile devices. In addition, the data collected is well understood and collection methodologies can be scaled up with minimum resources. That being said, it is also well established that urban sensing methods using GPS of mobile devices has problems of enhanced risk of breach of privacy when executed passively and need explicit user engagement when executed actively.

### 2.3.6 Wi-Fi

Wi-Fi is a wireless network connection protocol standardised by IEEE [2016]. It is a distributed server-client based system where the client connects to access points (AP). Every device in the network has a unique hardware specific MAC address, which is transmitted between the device and AP before the connection is made. The key feature of Wi-Fi infrastructure is that the network is distributed and the APs can be set up and operated by anyone locally unlike mobile networks. Since they are primarily used for Internet service provision, the protocol has priority for continuity of connectivity so the devices constantly scan for new and better connections. This is done through a probe request, which is detailed in later sections. With this background we can see that Wi-Fi provides a fair middle ground between an entirely network driven approach such as cellular network to an entirely user driven approach such as GPS. Since the network infrastructure is distributed and deployed for Internet it offers near complete coverage, is very resilient, and can encapsulate and reinforce civic space in cities<sup>17</sup>.

<sup>17</sup> Paul M. Torrens. Wi-fi geographies. *Annals of the Association of American Geographers*, 98(1):59–84, 2008. doi: 10.1080/00045600701734133

Though Wi-Fi is a location less technology, there are reliable methods to triangulate the location of the device by the signal strength and the known locations of APs [He et al., 2003, Moore et al., 2004, LaMarca et al., 2005, Dinesh et al., 2017, Lin and Huang, 2018]. This can overcome the usual shortcoming of GPS, which struggles for precision and accuracy in indoor and densely built environments [Zarimpas et al., 2006, Kawaguchi, 2009, Xi et al., 2010]. Utilising this, we can easily and quickly estimate trajectories of the mobile devices just using the Wi-Fi communication the device has with multiple known APs [Xu et al., 2013]. This can be used similar to the GPS trajectories to understand individual travel patterns [Kim et al., 2006, Rekimoto et al., 2007, Sapiezynski et al., 2015], crowd behaviour [Abedi et al., 2013, Mowafi et al., 2013, Shu et al., 2017], vehicular [Lu et al., 2010] and pedestrian movement [Xu et al., 2013, Fukuzaki et al., 2014, Wang et al., 2016, Taylor et al., 2019]. It can also be used in transportation planning and management to estimate travel time [Musa and Eriksson, 2011, Håkégård et al., 2018] and real time traffic monitoring [Abbott-Jard et al., 2013a].

Being a general network protocol designed to be used by mobile devices, Wi-Fi devices relay a range of public signals known as probe request frames on regular intervals throughout its operation, for the purpose of connecting and maintaining a reliable and secure connection for the mobile device [Freudiger, 2015]. These signals can be captured using inexpensive customised hardware,

non-intrusively and in turn to be used for numerous applications. In addition to a uniquely identifiable MAC address, these signals include a range of other information which when combined with the temporal signatures of the signals received can help us understand the nature and identify the devices which are generating these signals. These device/user fingerprinting techniques are demonstrated by [Franklin et al., 2006] and [Pang et al., 2007b] and the unique MAC addresses and associated information can successfully track people across access points <sup>18</sup>, their trajectories [Musa and Eriksson, 2012], the relationship between them [Cheng et al., 2012, Barbera et al., 2013, Cunche et al., 2014] and predict which of them will be most likely to meet again [Cunche et al., 2012]. Using the semantic information present in these probe requests it is possible to understand the nature of these users at a large scale [Di Luzio et al., 2016]. Using the received signal strengths from pre placed devices we can monitor the presence and movement of entities that are not even carrying a Wi-Fi enabled device [Elgohary, 2013].

Because of the security and privacy risks posed by the Wi-Fi protocol's use of hardware based MAC address, various methods to strengthen the security have been proposed [Pang et al., 2007b, Greenstein et al., 2008]. The randomisation of MAC addresses has become more mainstream in mobile devices with the introduction of it as a default operating system behaviour in iOS 8 by Apple Inc. Since MAC randomisation is not a perfect solution [Mathieu Cunche, 2016] there have been numerous attempts to fingerprint unique devices from the randomised anonymous information present in the probe request frames for the purposes of trajectory tracking and access point security. The methods used are decomposition of OUIs where detailed device model information is estimated by analysing an already known dataset of OUIs [Martin et al., 2016]; Scrambler attack where a small part of the physical layer specification for Wi-Fi is used [Bloessl et al., 2015]; and finally, the timing attack where the packet sequence information present in the probe request frame is used [Matte et al., 2016, Cheng and Wang, 2016]. A combination of these methodologies has been proven to produce de-anonymised unique device information from randomised MAC addresses [Vanhoef et al., 2016]. In addition to tracking, Wi-Fi probe requests can be aggregated to uncover the urban wireless landscape [Rose and Welsh, 2010] and used to reveal human activity at large scales [Qin et al., 2013], pedestrian numbers in crowds [Schauer et al., 2014, Fukuzaki et al., 2015] and also counting people in hyper local scales such as queues [Wang et al., 2013]. With enough infrastructure we can aim to generate a real-time census of the city [Kontokosta and Johnson, 2016] and also predict the amount of time a device will spend around the sensor as well [Manweiler et al., 2013]. Similar to GPS data this can be used as an additional control layer for interpolation techniques such as map merging [Erinc et al., 2013]. A comparison of various approaches was done by <sup>19</sup> where through experiments on a telecom operator dataset, it was showed that using network-driven mobile phone location data is more advantageous compared to the widely used event-driven ones.

<sup>18</sup> Mathieu Cunche. I know your mac address: targeted tracking of individual using wi-fi, 2014. ISSN 22638733

<sup>19</sup> Fabio Pinelli, Giusy Di Lorenzo, and Francesco Calabrese. Comparing urban sensing applications using event and network-driven mobile phone location data. In *Mobile Data Management (MDM), 2015 16th IEEE International Conference on*, volume 1, pages 219–226. IEEE, 2015

<sup>20</sup> Chen Zhong, Michael Batty, Ed Manley, Jiaqiu Wang, Zijia Wang, Feng Chen, and Gerhard Schmitt. Variability in regularity: Mining temporal mobility patterns in london, singapore and beijing using smart-card data. *PLoS ONE*, 11(2), 2016. ISSN 19326203. doi: 10.1371/journal.pone.0149222

### 2.3.7 Consumer data

In addition to the direct data from the sensors themselves the content generated from the mobile devices such as social media data or smart-cards <sup>20</sup> can provide a viable proxy for estimating the level and nature of human activity. The use of geo-located tweets on the study of small-area dynamic population estimation [Ordonez and Erath, 2012, Marchetti et al., 2015, McKenzie et al., 2015, Lansley and Longley, 2016b], geo-demographics [Bawa-Cavia, 2011, Longley et al., 2015, Lansley and Longley, 2016a] and global mobility [Hawelka et al., 2014] has been thoroughly explored. These data sources are shown to be useful in social sciences [Crane and Sornette, 2008], abnormal event detection [Chae et al., 2012] and analysing urban environments [Sagl et al., 2012]. It can also be used as a control layer for interpolation techniques we discussed earlier [Lin and Cromley, 2015].

## 2.4 Research Gaps and Opportunities

In this section we summarise the previous sections to find out the best possible technology for further research and discuss the research gaps and opportunities available to us. Table 2.1 summarises the above discussion to evaluate all the relevant technologies that can be used for the data collection and analysis for the study of human activity at a granular level.

Table 2.1: Evaluation of different technologies or approaches that can be used for data collection.  
*\*independence from secondary data collected by a third party.*

Technology	Interpolation	Bespoke	Cellular	GPS	Wi-Fi
Coverage	Local	City	All	Local	All
Certainty	Very Low	High	Medium	High	Medium
Independence*	Low	Very High	Low	Medium	High
Intrusiveness	Low	Medium	High	High	Medium
Granularity	Very Low	Very High	Medium	High	High
Ease of Collection	Medium	Low	Medium	Low	High
Scalability	Medium	Low	High	Medium	High
Privacy Risk	Low	Medium	High	High	Medium

We can observe that Wi-Fi offers the best possible technology in terms of flexibility and scalability for data collection through mobile devices at an individual level while posing some risk to privacy of participants and involves uncertainty regarding the field of measurement.

From the literature search we can summarise that there is a considerable opportunity in the collection and analysis of mobile phone based data for measuring hyper-local, spatio-temporal dynamics of human activity. The potential for research gaps broadly fall under the following areas,

### 2.4.1 Ambient population analysis

Previous research in this area of study has been limited to either national/ regional level studies using centrally collected night time population data such as censuses or to area level studies conducted with mobile devices based technologies. Though there were some efforts in collecting and using mobile phone data at national/ regional level we have never been presented with

**Opportunity 1:** Design and collection of national/regional, longitudinal, grass root level data set which enables study of population both spatially and temporally.

such unprecedented level of data available now. The explosion of consumer data both publicly available and privately held presents previously unseen opportunity and also limited by the privacy concerns that arise with them. There is an immense opportunity to collect and standardise a large national level dataset which closely linked to the population distribution and movement in an anonymised way which then can be used to understand the distribution of population and its change. There is a need to extend such effort longitudinally which can give us insights into the change of such phenomenon in time. This has the potential to enable us to ask broad questions such as,

- What are the trends in the footfall in UK?
- What are the daily rhythm of different cities?
- How much a weather event affect economy of a region?

Such dataset, in conjunction with other consumer data sources, in addition to augmenting each other to improve their quality, can vastly improve our understanding of the structure and dynamics of population.

#### *2.4.2 Device fingerprinting*

The privacy concerns about the data collected from personal mobile devices has pushed the industry and users to find ways to anonymise the data generate over the last decade. All the mobility studies recording user trajectories across space and time are rendered infeasible with the cryptographic hashing and randomisation techniques employed by the devices. This along with progressive legislation such as General Data Protection Regulation have severely constrained the data available for mobility research. As we see later, even the estimation of ambient population is limited by these developments. Though there are prior research in this area, most are conducted from security perspective evaluating the robustness of the randomisation/obfuscation procedure. These research focus around de-anonymising the obfuscated data to recreate the personal information from them thus demonstrating vulnerabilities and risks for the users. In this context, there is a clear gap for research in to methods to rather carry out fingerprinting of these devices using patterns in the data to create useful information from them without actually de-anonymise the data. This can lead to production of data-sets and methodologies which will enable us to,

- Get accurate estimation of ambient populations.
- Understand the movement of the population in space and time.

*Opportunity 2: Developing models and methods to identify anomalies in the data and underlying events causing them*

#### *2.4.3 Event Detection*

Having granular spatio-temporal data on population at an area level also enables us to look at the activity of people at this scale. For example, the spike in Wi-Fi activity at a certain area at a certain time can illuminate us with a specific event that is happening in that area. Thought research have been done on this area using social media data, a longitudinal data-set collected using mobile technology can enable us to formalise the models needed to identify anomalies, quantify the causation of such anomalies to real world event. The

*Opportunity 3: Developing models and methods to identify anomalies in the data and underlying events causing them*

near real-time aspect of such research also provides us opportunities in fields of disaster management, smart cities etc. There are opportunities to ask questions such as,

- How did the tube strike affected London?
- What were the hot spots for New years celebration?
- What was effect of a road closure in specific part of the city?

#### 2.4.4 *Pedestrian Flow*

Similar to the device fingerprinting, estimating and understanding pedestrian flow in the street network has immense opportunities since the anonymisation of mobile devices has taken off. Even when the problem of identifying unique fingerprints of users in the data has not been solved, there is a need to understand the overall performance of the street network in terms of pedestrian flow just from the precise, granular data available, especially when the data source is as unstructured and noisy as the Wi-Fi sensors. This problem can be approached in two ways,

1. Probabilistic approach - Where the relationship between the temporal change in volumes at locations are modelled. For e.g. how much and how often the footfall counts at one location mirrors/ follows other location gives us an idea of how many pedestrians move from one location to the other.
2. Interpolation - Where the relationship between the locations are defined in terms of multiple variables such as how similar they are, how close they are etc. These relationships can in turn used to build a graph of locations and use this graph as a source to interpolate other locations.

#### 2.4.5 *Nature and Change of Places*

Though there are extensive research in using ambient population and people's movement to understand the form and function of the space, the mobile technologies have introduced the opportunity to remove the subjectivity from them. With access to granular and long-term data sets, we can hope to look into the how the places have changed over time and how the external factors such as policy and economy has affected them. There are opportunities to ask questions such as,

- How does UK's exit from EU affect its high streets?
- Has a specific area has become more or less vibrant?

*Opportunity 4: Estimating flow of pedestrians in the street network from Wi-Fi data*

*Opportunity 5: Using long term data to detect the nature and change of form and function of a place.*

# 3

## *Collecting Wi-Fi Data*

From the literature review in Chapter 2, we observed that out of all the technologies discussed, Wi-Fi seems to be the most promising one. We observed the advantages of Wi-Fi based data collection as,

- Universality as a standard technology globally,
- Independence from other types of data sources or infrastructure,
- High level of granularity both spatially and temporally,
- Possibility of passive data collection,
- Extreme ease of collection in terms of cost and effort and
- Scalability to cover study large areas.

Though it has its pitfalls in terms of intrusiveness resulting in risk to the privacy of the users, bias and uncertainty, Wi-Fi provides us with a strong base framework for fulfilling the opportunity to design and collect a large, long term and granular dataset which can be used for studying human activity.

In this chapter, we continue our research by looking at Wi-Fi technology closely to understand how it can be used to achieve the aforementioned goal. We start by looking at the Wi-Fi specification <sup>1</sup> and focus on the information available within the Wi-Fi probe requests. We then design and implement a series of data collection exercises which collect probe requests in various location with increasing level of complexity for analysis. We explore these datasets briefly to understand the usefulness of each set of information present in the probe requests along with the uncertainties in them. We also introduce ‘Smart Street Sensor’ project - a national scale effort for collecting Wi-Fi data at high streets across United Kingdom. Finally we summarise the data collection procedure with a detailed look at the each set of uncertainty in these datasets and draw conclusions on further lines of research into alleviating the uncertainty and noise so that they can be used to estimate human activity with confidence.

### *3.1 Wi-Fi as a Source of Data*

Since the formation of ‘Wi-Fi alliance’ in 1999 to hold the trademark, Wi-Fi (Wireless Fidelity) has become synonymous to the IEEE 802.11 standard based internet connectivity. Today almost all devices use this standard to create and connect to local area networks wirelessly. Due to its high fidelity and immense throughput up to 1 Gigabits per second, Wi-Fi has become the

<sup>1</sup> IEEE. IEEE standard for information technology-telecommunications and information exchange between systems local and metropolitan area networks-specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. *IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)*, pages 1–3534, Dec 2016

choice of technology for wirelessly transferring large amount data through networks. The adoption of ‘smart’ mobile devices - Smartphones, across the world has further cemented its position as one of the most ubiquitous technology which every one uses every day. In developed economies such as UK this has never been more true and having an infrastructure to serve and receive Wi-Fi signal greatly affects the ability to connect to internet in many areas. With close to 87%<sup>2</sup> of the population carrying one or more of these smart devices with Wi-Fi capability, provision of Wi-Fi as a service has become essential for any place making Wi-Fi one of the most used technology to access internet along with mobile networks.

Though the end goal of internet connectivity is the same, Wi-Fi greatly differs from internet connectivity through mobile networks such 3G/4G. The first difference is the range of the network, unlike mobile infrastructure where single tower can serve mobile phones for miles, Wi-Fi is designed to be an extension of the wired networking creating short range network with a range up to 20 meters. Due to this low-range and high throughput property, Wi-Fi is used primarily as a distributed infrastructure operated by owners of premises as a means to provide high speed connectivity to the users of these buildings unlike the large, national level, monolithic infrastructure that runs the mobile network. This creates a situation where urban areas are populated by hundreds and thousands of these small area networks to which any mobile device can connect to. Unlike the mobile service providers and its customers, these Wi-Fi networks and mobile devices don’t trust each other with specialised hardware. This creates a need for an introduction procedure - a sort of handshake between them where they exchange information about themselves. Moreover since these mobile devices constantly move across these Wi-Fi networks, it became necessary for them to carry out these ‘handshake’ process regularly and frequently so that they can traverse between the networks without loss of connectivity. This need for constant look out for new networks is solved by the ‘Probe requests’.

### 3.1.1 *Probe requests*

There are numerous iterations and versions of the IEEE 802.11 standards but essentially all of them operate by exchanging packets of information called ‘datagrams’ or ‘frames’. These frames have the information that is being exchanged along with the meta data about the information and information on the device that is sending them. Some of these frames have special purposes and one such purpose is the ‘network discovery’. The frames used for this purpose by the mobile device and the Access Point (AP) are called the ‘probe request’ and ‘probe response’ respectively. Though the actual information exchanged between these devices are usually encrypted, these probe requests are unencrypted and are accessible to any device which is listening. The structure of a probe request is shown in Figure 3.1.

We can observe that the fixed part of the frames which has to be present in all of them are in the front showing the identity of the mobile device generating the frame along with one which is receiving the frame. There are additional meta data such as sequence number of the frame, controls denoting where the frame starts and ends. There is also a dearth of variable information

<sup>2</sup> Deloitte. Mobile consumer survey - united kingdom, 2018. URL <https://www.deloitte.co.uk/mobileuk/>

which can be used to transfer data. For probe requests the destination device is set as ‘broadcast’ and the variable part usually contains the payload but for probe request frames, this contains ‘information elements’ which has data regarding the capabilities of the device organised in units known as ‘tags’ or ‘parameter sets’. The significant information present in a probe request is detailed in Table 3.1 and full list of information available from a probe request is shown in the form of a sample probe request in appendix 7.2

Field	Notes
Source Address	Media Access Control (MAC) address
Time stamp	Precise time at which the frame is received
Signal Strength (RSSI)	The strength of the received signal
Length of the frame	Total length of the frame in bytes
Duration of transmission	Time it took to transmit the frame in milliseconds
Information Elements	List of various information about the device
Known SSIDs	Networks that are already known to the device

Table 3.1: Significant information included in a probe request

Essentially the above information is sent over and over by the mobile device expecting a reply from the APs around it so that it can keep a list of networks it can connect to. This process is usually carried out even when the Wi-Fi is switched off in the operating system so that the connection times are faster once it is switched on. Moreover operating systems use the replies they get for these probe requests and triangulate the device location with respect to the APs with known location information thus acting as a quick and easy localisation solution which along with the above makes this probing process almost non-stop.

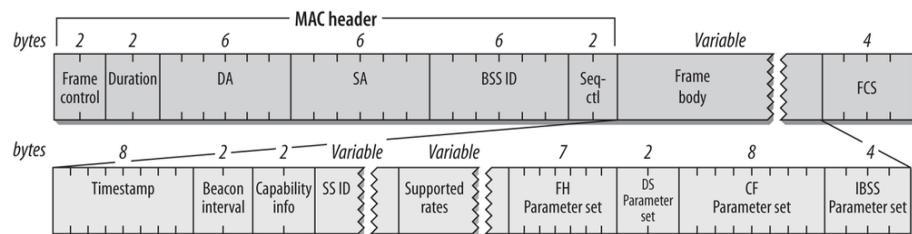


Figure 3.1: Structure of a probe request frame. Source: IEEE 802.11 specification

### 3.1.2 MAC address

Media Access Control (MAC) address is a 12bit unique identifier assigned to a device in a network, it is similar to the Internet Protocol (IP) address but is assigned at the interface controller level and are usually assigned by the manufacturer of the device. Although the IP address of a mobile device might change regularly, the MAC address may remain just the same even for its life time making it akin to a unique identifier of a device and much more significant. It has two parts the first 6bits are known as Organisationally Unique Identifier (OUI) and gives us information about the manufacturer of the network card. Organisations need to register with IEEE to be assigned an OUI which they can use to generate a full MAC address. The second 6bit is the device part of the MAC address which is unique for the device from the manufacturer. Together both form the full MAC address which is unique to every device globally.

The biggest draw for using Wi-Fi for mobility analysis comes from the fact that this globally unique ID is sent out regularly by mobile devices and can be collected passively through probe requests. As we saw in our literature review, this also creates an immense risk in terms of infringement of privacy both for the manufacturer and the user. Manufacturers of critical hardware components who do not want their unique MAC address to be publicised usually opt for registering a ‘private’ OUI which will not be given out to other manufacturers and will not be published publicly. Users (their mobile devices) who don’t want to be tracked using their MAC addresses use a temporary MAC address which is unique only to the local network - ‘local’ OUI rather than using a ‘global’ OUI for unencrypted communications and switch to their original MAC address when a trusted encrypted connect has been established. This lack of uniqueness can be inferred from the second character of the MAC address being E, A, 2 or 6. Though this provides reasonably better privacy to the mobile users it also limits our ability to just use the MAC address from the probe requests as previous studies conducted with Wi-Fi.

To summarise, MAC addresses are two types based on whether it has a public OUI or a private OUI. This distinction does not affect their uniqueness or usefulness in mobility research but hinders us from knowing about the device from the MAC address. MAC addresses are also of two types based on them being unique globally or just in the local network. This distinction affects the feasibility of using them for device tracking or for studying movement of the users.

### *3.2 Initial Experiments*

With our theoretical understanding of the Wi-Fi standard and its capabilities, we move on to looking at the Wi-Fi landscape in real-world. We achieve this by designing small independent experiments where we record the Wi-Fi probe requests within controlled conditions along with the knowledge of the ambient population of the field of measurement. We then look at the collected probe requests, examine them in detail to look at their properties, aggregate them to footfall counts and compare them with the real-world counts to get an overall idea of how well they translate into real counts. The aim of these experiments to know more about the probe requests data and pick out the uncertainties and opportunities present in them. The objectives here are,

1. Design a simple method to collect probe requests.
2. Select locations with different levels of complexity.
3. Collect real-world data through manual counting.
4. Analyse the probe requests to extract useful information.

#### *3.2.1 Experiment Design*

The first step was to design a simple method to collect Wi-Fi probe requests. We accomplish by utilising the application - *tshark*<sup>3</sup> on a regular laptop. First we put the Wi-Fi module of the laptop in ‘Monitor mode’ where it behaves as a wireless access point. Then we run tshark to collect the data in CSV format using the following command under a shell.

<sup>3</sup> Gerald Combs and Contributors. Wireshark - network protocol analyzer. <https://www.wireshark.org/about.html>, 2018

```

1  #! /bin/bash
2  tshark \
3    -I -i en0 \
4    -T fields \
5    -E separator=, \
6    -E quote=d \
7    -e frame.time \
8    -e frame.len \
9    -e wlan_radio.signal_dBm \
10   -e wlan_radio.duration \
11   -e wlan.sa_resolved \
12   -e wlan.seq \
13   -e wlan.tag.length \
14   -e wlan.ssid \
15   type mgt subtype probe-req and broadcast

```

The fields marked with `-e` correspond to time stamp when the packet was received, total length of the packet, reported signal strength, duration for which the packet was transmitted, the MAC address of the source device, sequence number of the packet assigned by the source device, length of the tags attached to the packet and the network for which the probe request is being sent for. The manufacturer information is extracted from the `wlan.sa_resolved` field into a separate column and the original field is hashed using the SHA256 algorithm implemented in R. In addition to this, the pedestrians next to the sensor were counted manually by the surveyor.

### 3.2.2 Living room

The first set of experiment was conducted with the laptop in the researcher's living room. The room was known to have 2 phones and an Android TV device and the whole house had an occupancy of 3. Probe requests were collected on 15 Nov 2015 from 19:44 to 21:15 with a gap of 15 minutes in between. We collected around 3000 probe requests at the rate of 38 requests per minute. The total number of probe requests collected every minute is shown in Figure 3.2.

The first thing we looked to convert this to footfall is the mac address along with OUIs. Initial thoughts on this. We found that it is not enough. Strong evidence that mac address randomisation is affecting the numbers.

Other information that can help here - Signal strength, tags, SSID, length, sequence numbers and duration. Signal strength shows promise! Length seems useful as it give uniqueness! Duration and length are directly related and gives no extra information. SSID and tags have very low or varied data they have no use. Sequence numbers are interesting and shows most promise. We need to see the variation and usefulness of length, so need more data.

### 3.2.3 University Hall

This was conducted to examine the previous results in a larger set of data in a real world setting. Conducted in one of the main corridors - Southern cloisters of UCL with lot of pedestrian traffic. There were also seating areas

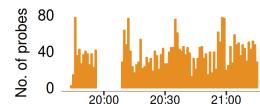


Figure 3.2: Number of probe requests collected every minute on 15 October 2017

across the corridor where students work with their computers. The area is also used heavily for lunch for large contingent of visitors. Collected around 14750 probe requests collected and 652 users were counted walking around the sensors manually.

Major conclusion is that confirmation that Signal strength is still useful. Length information is not that useful in certain standardised manufacturers.

### *3.2.4 Oxford Street*

Aim is to validate the filtering and clustering methodology against the scale and complexity of data collected in an open public area such as a retail high street. We also aimed to find the algorithm which was best suited for the classification of signal strengths as 'low' and 'high' in order to filter out the background noise. The data was collected at Oxford Street, London on 20 December 2017 from 12:30 to 13:00 hrs, Wi-Fi probe requests were collected using the sensor described in Section and pedestrian footfall was manually recorded using the Android app - Clicker bala2018clicker. Being located at one of the busiest retail locations in the United Kingdom, the Wi-Fi sensor captured approximately 60,000 probe requests during the half hour period; 3,722 people were manually recorded walking on the pavement during that time. The surveyor positioned himself at the front of a store while carrying the sensor in a backpack and counted people walking by the store on the pavement (3m wide approximately) using a mobile phone. The sensor was kept as close to the store window as possible, and the manual count was done as a cordon count in front of the store.

The analysis and use of this dataset is

### *3.2.5 Summary and Discussion*

## *3.3 Pilot Study*

The ai

### *3.3.1 Methodology*

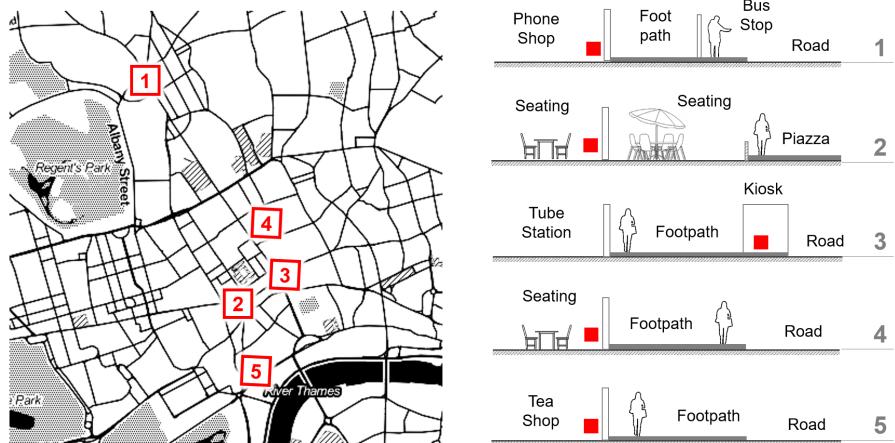
The methodology set out above was implemented in five different Central London locations at different times. % Sensors were installed and data collected for extended periods of time. Sensors were installed and data collected for extended periods of time.

We then applied the methodologies discussed earlier to arrive at estimated pedestrian footfall and compared them with the corresponding manual counts. We finally evaluated the effectiveness of the processes with the Mean Absolute Percentage Error (MAPE) at the locations and report our findings below.

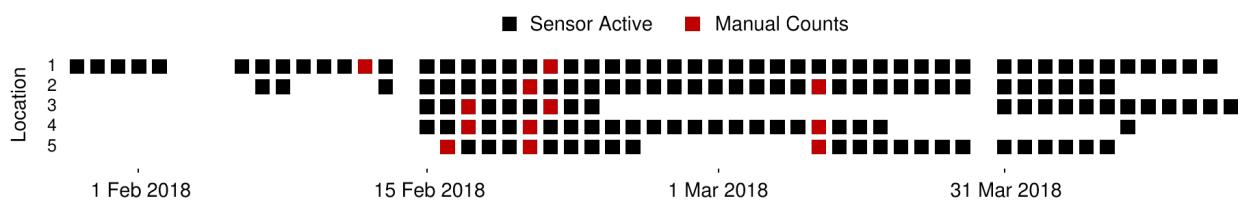
### *3.3.2 Locations*

Locations where sensors were installed, volume and speed of probe requests collected by the sensor and total pedestrians manually counted. The data occupies around 1.8 GB on disk when encoded in text format. The locations at which the data were collected are shown in Table . The locations were

Figure 3.3: Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data



chosen for their diverse site conditions and unique sources of noise around the potential location of the sensors. The position of the sensor at these locations with respect to the context is shown in the Figure. We can see that Location 5 is the 'cleanest' with one clear stationary source of noise (phone shop) while location 2 is the most complex due to the proximity of seating areas to the sensor. The sensors were operational throughout February and March, while manual counts were conducted in these locations in half hour sessions on at least two different days. For the purposes of comparing with ground truth, we considered the data from sensors which correspond to the 12 sets of available manual counts. The schedule of data collection is shown in Figure .



ID	Location	Type	Installation notes	Probe Requests	Counted Footfall
1	Camden High Street	Phone Shop	Bus stop in front	9.9 (297)	3683 (33)
2	Central St.Giles	Restaurant	Seating area on both sides	3.9 (169)	0346 (05)
3	Holborn Station	Info. Kiosk	Overlooks station entrance	4.3 (303)	2956 (46)
4	Brunswick Center	Fast Food	Has seating area on one side	3.4 (210)	0960 (12)
5	The Strand	Tea Shop	Phone shop next door	8.4 (382)	1969 (21)

Figure 3.4: Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data installed sensors

### 3.3.3 *Data Collection*

## 3.4 *Smart Street Sensor Project*

The Smart Street Sensor project is one of the most comprehensive study carried out on consumer volume and characteristics in retail areas across UK. The project has been organised as a collaboration between Local Data Company (LDC) and Consumer Data Research Centre, University College London (CDRC, UCL). The data for the study is generated independently within the project through sensors installed at around 1000 locations across UK. When completed, the project will serve as the first and unique comprehensive research into the patterns of retail activity in UK high streets.

The primary aim of the project is to improve our understanding of the dynamics of the high street retail in UK. As we saw in our literature search, unlike online retail, this involves quantification and measurement of human activity at small scales, such as high streets which already the subject of active research. The key challenge in this area is the collection of data at smallest scales possible with minimal resources while not infringing on people's privacy. This challenge when solved can provide immense value to occupiers, landlords, local authorities, investors and consumers within the retail industry. The project aims to facilitate decision making by stakeholders in addition to the tremendous opportunities for academic research.

### 3.4.1 *Methodology*

As a first step, various locations for the study were identified by CDRC to include a wide geographical spread, different demographic characteristics and range of retail centre profiles. A custom footfall counting technology using WiFi based sensors was also designed, developed by LDC and the sensors were installed the identified locations. The sensor monitors and records signals sent by WiFi enabled mobile devices present in its range. In addition, the number of people walking by the sensor was counted manually for short time periods during the installation. The project aims to combine these two sets of data to use as a proxy for estimating footfall at these locations. The potentially identifiable information collected on the mobile devices is hashed at sensor level and the data is sent to central server via encrypted channel for storage. This data is then retrieved securely for the preparation of the commercial dashboards by LDC and for research purposes by CDRC users. The project began on July 2015 with the first sensor installation and has grown to an average of 450 daily active sensors as of January 2017.

The data is collected through set of SmartStreetSensors (shown in Figure 3.1), a WiFi based sensor which when installed acts as a WiFi access point and collects specific type of packets (probe requests) relayed by mobile devices which are within the device's signal range and are searching for available access points. The sensor is usually installed on partnering retailer's shop windows so that its range covers the pavement in front of the shops. The installation and calibration of device with respect to the shop window and the pavement is illustrated in Figure 3.2. There is also a small percentage (3%) of the devices which are installed within large shops to monitor internal footfall. Each device collects data independently and uploads the collected

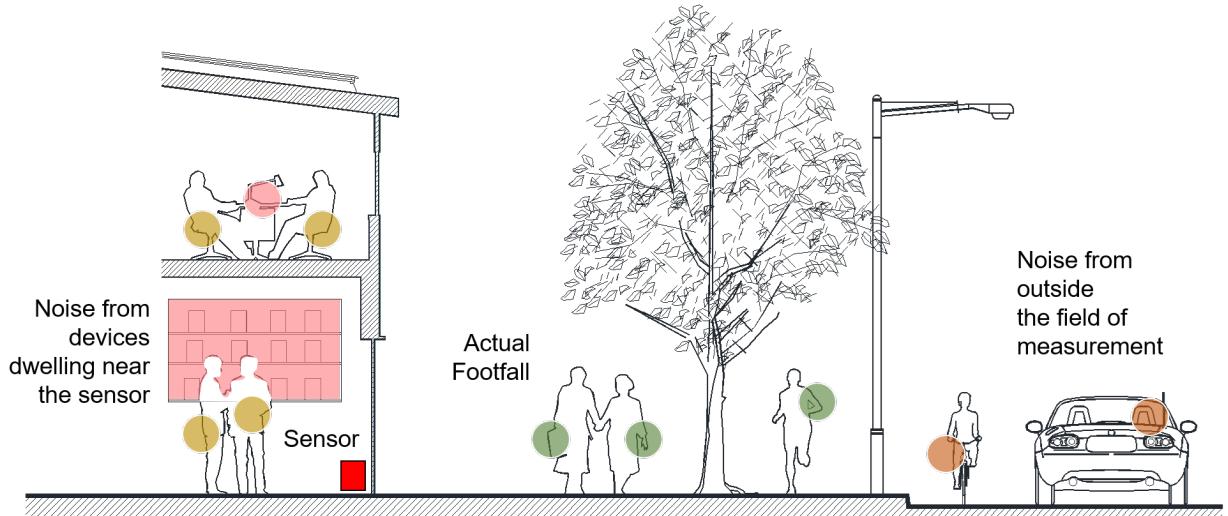


Figure 3.5: Outline of the 'Medium data toolkit' de-

vised to collect process data to a central container at regular interval 5 minutes through a dedicated 3G mobile data connection. The sensor hardware has been improved over the course of the project and currently has built in failure prevention mechanisms such as, backup battery for power failures, automatic reboot capabilities and in-device memory for holding data when internet is not available. The hardware versions and the corresponding features are detailed in Table 3.1.

smartstreetsensor - figure. system architecture - figure.

### 3.4.2 Locations

locations - map locations - table

### 3.4.3 Data Description

data description - table. sample probe request - code. installation of sensors - figure. descriptive statistics - table.

## 3.5 Discussion

In the previous sections we designed and implemented data collection processes to arrive at 3 different datasets - Small experiments, Pilot Study and Smart Street Sensor project. The small experiments were designed as way to collect as much data as possible from the probe requests for short periods of time aiming to collect small sets of comprehensive data under controlled conditions for exploratory purposes. The pilot study extended this further by collecting data for a longer time in real world conditions aiming to validate the insights collected with the experiments and the methodologies we devise for the research. The Smart Street Sensor is the most comprehensive study which collects very small focussed set of data at a national level for very long periods of time. These datasets give us a well rounded set of data to set up our toolkit and devise out methodologies. The summary of the datasets in terms of their characteristics is show in table 3.3.

\*approximate

Dataset	Locations	Time	Detail	Purpose
Small Experiments	3	30 - 60 mins	High	Exploratory analysis
Pilot Study	5	6 weeks*	Medium	Devising and calibrating methodologies
Smart Street Sensor	1000*	4 years*	Low	Real world insights

Table 3.3: Summary of the collected data-sets.

Before we move on to develop methods to process the data into information on footfall in these locations, the crucial action is to look at the possible biases and uncertainties in these datasets arising due to the data collection methodology and from the broader context. These form the framework on which built our data processing pipeline where we propose to solve each of these uncertainties in each step.

From our understanding of the data, we observe that the major sources of uncertainties are regarding the range of the sensor, the frequency at which mobile devices generate probe requests, the way and rate at which the mobile devices randomise their MAC addresses, the collisions caused due to the hashing of the MAC addresses and finally the gaps introduced by the failure of the sensors. There is an inherent bias to these data caused by the mobile phone ownership in the population which varies across time, location and demography. We discuss each of these uncertainties and biases in detail below,

### 3.5.1 Range of the sensor

The first and foremost uncertainty we face with wireless sensors such as Wi-Fi and Bluetooth is the delineation of the field of view of the sensor. Though the Wi-Fi signals can be partly managed or restricted by manipulating their strength, there is no reliable way to precisely delineate a survey area for these sensors. The manipulation of signal strength can be done by installing metal shields around the sensors to block certain directions and prioritise others but the method cannot block out all the signals and will leave some uncertainty about where the probe requests are coming from. Moreover, strength of the signal received from a mobile device by the Wi-Fi access point depends on numerous factors such as,

1. Distance between the mobile device and the Access Point.
2. Thickness of the objects present in between them.
3. Nature of obstructions such for e.g. metal vs glass
4. Interference from other wireless devices.
5. Power level of the transponder of the Access Point.
6. Power level of the transponder of the Mobile device.
7. Even atmospheric conditions such as humidity, temperature etc.

The signal strength drops non-linearly when moving away from the access point as shown in Figure 3.6 and there is a *close-range non-monotonicity* as well - where within 10 feet, A closer device can report lower signal strength than a farther one [Cisco, 2008]. The relationship between the two is given by the equation <sup>4</sup>,

<sup>4</sup> Zengzeng, Andy, and Cabral. Calculate distance from rssi. URL <https://bit.ly/20hskf9>

$$\log_{10} d = \frac{(P_o - F_m - P_r - (10 \times n \times \log_{10} f) + (30 \times n - 32.44))}{10 \times n} \quad (3.1)$$

Where,

$d$  = distance - Sensitivity of the receiver

$F_m$  = Fade Margin - Sensitivity of the receiver

$n$  = Path-Loss Exponent, ranges from 2.7 to 4.3

$P_o$  = Signal power (dBm) at zero distance - Measured by testing

$P_r$  = Signal power (dBm) at distance - Measured by testing

$f$  = signal frequency in MHz - Specific to the hardware

All these factors vary widely in real world conditions at each location depending on where and how the sensors are installed. They also vary widely over time due to changes in the context and vary across different directions at each location as well. This makes it extremely difficult to model the distance between mobile device and the access point as a function of the received signal strength. The equation 3.1 can be approximated and simplified as,

$$R = (-10 \times \log_{10} d) + A \quad (3.2)$$

Where  $R$  is the reported signal strength and  $A$  is the signal strength at 1 meter. Though equation 3.2 can help us to infer the distance of the mobile device roughly, the location wise uncertainty of this method makes it lose the meaning when compared across locations.

From the above we can conclude that it is almost impossible to delineate the field of measurement precisely and accurately by simple methods using the information present in the probe requests. This leads to uncertainty in the data collected which needs to be resolved with explicit assumptions or specific methods to reduce the resulting noise. We require a method to isolate this noise from data generated by devices within the field of measurement. This method needs to be independent of the micro site configuration, temporal changes in the context.

### 3.5.2 Probe request frequency

The second uncertainty we face is the frequency at which mobile devices generate the probe requests which varies wildly. The number of probe requests generated from a mobile device depends on,

1. Manufacturer of the device. E.g. Samsung vs Apple
2. Version of the software running on the device. E.g iOS 7 vs iOS 8
3. State of the device. E.g. If it is already connected to internet? Has the location services been switched off?
4. The number of access points already known to the device.

Studies done by Freudiger [2015]<sup>5</sup> has shown that the number of probe requests generated by a mobile device vary widely across manufacturers such as Samsung, Apple and LG, across different states the devices are in such as charging, screen being on, Airplane mode being on etc and depends heavily on the number of access points known to the device. It is also seen from our initial experiments that these probe requests are generated in short bursts rather than being generated at regular intervals. This makes predicting a base factor for calculating a number of mobile devices based on the number of

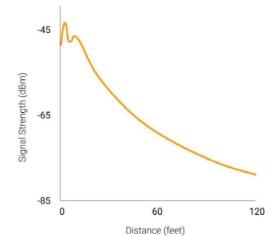


Figure 3.6: The decay of signal strength (RSSI) with respect to distance.

Source: Wi-Fi Location-Based Services, Cisco

<sup>5</sup> Julien Freudiger. How talkative is your mobile device?: An experimental study of Wi-Fi probe requests. In *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, WiSec '15, pages 8:1–8:6, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3623-9. doi: 10.1145/2766498.2766517. URL <http://doi.acm.org/10.1145/2766498.2766517>

probe requests received much more complex. The variety of device models available and the pace of change in software that run these models further complicates this. Though we can simply aggregate these probe requests based on the unique information on them, in absence of such information it becomes extremely critical. We to consider this uncertainty in detail while making any simple assumptions on the relationship between number of probe requests and the number of mobile devices that generated them.

### 3.5.3 MAC address randomisation

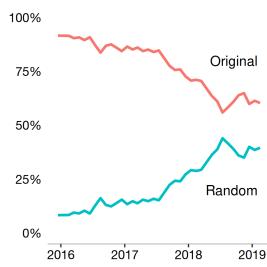


Figure 3.7: Increase in the share of randomised MAC addresses compared to non-randomised original ones over the years.

From data collected at Regent Street, Cambridge.

Randomisation of MAC address is one of the recent uncertainty introduced in the data. As we saw in section 3.1 MAC address is the unique identifier for each mobile device and we aggregate the footfall numbers based on this. Since the probe requests are transmitted unencrypted and can be received by any access point, this is one of the biggest leak of personal data which occurs in the Wi-Fi based communications. Modern mobile devices solve this problem by using a randomised MAC address for the probe requests which can result in large over estimation of number of mobile devices in the vicinity.

The method of randomisation and the frequency varies widely between device manufacturers and also changes as new versions of the software are released. This seriously affects the usefulness of the data long term where methods designed to overcome this randomisation can be rendered inefficient in the future. Figure 3.7 shows the increase in the share of randomised MAC addresses since 2015. We can observe that in addition to the overall upward trend there are bursts of increase around late 2016 and 2017 which coincides with the release of new mobile operating systems. This makes it necessary for devising a method to overcome MAC randomisation problem to be able to uniquely fingerprint devices so that they can be aggregated together. As we saw in our literature review, this is also one of the major opportunities in research on human mobility using Wi-Fi data.

### 3.5.4 Mobile Device Ownership

One of the major external bias in all the datasets collected from mobile devices is the overall volume and nature of the ownership of these devices. The ownership of mobile devices, specifically Wi-Fi enabled ones, have been on the rise since 2005. Though mobile ownership has reached unprecedented level in recent years, there is still an underlying increasing trend present in the ownership of these devices which manifests itself in the collected data. Moreover the mobile ownership varies widely between demography of age and geography as well. Figure 3.8 shows the mobile ownership across age groups in UK from 2012 to 2018. We can observe that the older age groups are under-represented in our data. This needs to be taken into consideration while using this data for extrapolating demographic conclusions out of it. In addition to this, the overall upward trend needs to be adjusted assuming 1% increase monthly and 0.2% weekly when using this data across long periods of time.

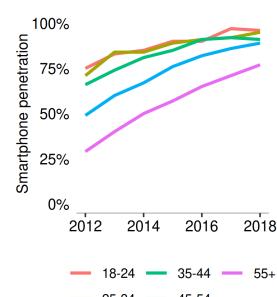


Figure 3.8: Smartphone penetration by age group in United Kingdom (2012-18)

### 3.5.5 Missing Data

Being collected by distributed set of sensors located in busy real-world scenarios, the data has a large number of gaps as well. These gaps are caused by various reasons such as, failure of the sensor hardware and software, failure in internet connectivity to send the data back, External factors such as store closures which can cause power loss, regular disruptions such as software update and maintenance and finally other factors such as unauthorised tampering and unplugging of the sensor. This leads to a dataset which contains several small and medium sized gaps as shown in Figure 4.2. More over, the Smart Street Sensor project is implemented and managed with commercial motive, the sensors are installed and uninstalled at locations as retail partners join and leave the project. This leads to an uneven availability of data across locations over longer time periods which creates challenges while aggregating data across locations. We need to implement a methodology to fill in these gaps which considers the periodic patterns in the data. We also need to devise a measure for aggregating the counts across locations which removes the bias introduced by long term gaps in data.

### 3.5.6 MAC address collisions

Finally, from the initial analysis we have observed that there are few instances of collisions occurring in the hashed MAC addresses. This has been observed as unique hashed MAC addresses appear at different locations within a short period of time which cannot be explained by the physical travel by the user between these locations. These collisions are caused by the limitation of the hashing algorithm used and exist only in very large amounts of data. It is important to note that these collisions are specific to non-randomised MAC addresses as we don't expect any consistency within the randomised ones. Even though this is an inevitable side effect of the hashing process, the probability of such occurrence is very low and is calculated as  $2^{-n}$ , where  $n$  is the number of bits in the output of the hashing algorithm. The total number of estimated collisions between  $m$  unique values is given by,  $2^{-n} \times \binom{m}{2}$ <sup>6</sup>. This translates to around 100 collisions across a million unique devices with a 32 bit algorithm and 2 collisions across 10 Billion devices when using a 64 bit algorithm. Though these collisions might cause issues in granular mobility models, for long term and broad studies where we don't track individual devices, they can be safely ignored.

<sup>6</sup> mikeazo and poncho. Formula for the number of expected collisions. URL <https://bit.ly/2Y56zYL>



4

## *Processing the Data into Footfall*



## 4.1 Data Toolkit

BIG DATA AND ITS ANALYTICS promises huge benefits in terms of value realisation, cost reduction, insights but it also introduces a numerous pitfalls<sup>1</sup>. With developments in information technology, mobile communications and internet of things, large assemblages of data are readily available leading to immense possibilities in research. But when we analyse these data at such scale, we also encounter a large amount of added complexity and cost. Hence it is important to be careful in choosing the methods and tools in dealing with big data where we should look to devise right methods and tools for the right problems. Moreover in several disciplines, such as statistics and geography etc., the existing methods and tools are already developed for dealing with large scale data. These methods along with improvements in hardware has made the processing big data in these disciplines possible without a major changes in workflow. In the current environment of constant change and growth of sources of data, we cannot afford to lose the opportunity to extract information from them while trying to create a perfect, future proof approach in dealing with them. We need to move fast with a pragmatic approach where we look at other disciplines and adopt best practices and solutions in them and develop consistent approach for our needs rather than reinventing the wheel.

In the previous chapters we looked at various methods we devised to collect and process data from Wi-Fi probe requests emitted by phones. Though we discussed the methods conceptually, we left out the rationale behind choosing the toolkit employed to implement those methods. In this section we elaborate the thought process and rationale behind these decisions. We start by discussing the concept of 'Big Data' in general and look at previous literature to understand its definition, nature and the challenges they pose. Then we look at the data-sets we collected through the pilot studies and the 'Smart Street Sensor' project and evaluate them in terms of the dimensions of the big data. We also discuss the challenges faced in dealing with our dataset in detail and try to understand the requirements for devising a toolkit for it. Finally we put together a toolkit to suit our datasets built from simple small UNIX tools.<sup>2</sup>

### 4.1.1 What is 'Big Data'?

With the proliferation of internet enabled personal devices, we have quickly moved from data sparse environment to a data rich one. We can even confidently say that we are in an age of data deluge where the amount of data which are collected and stored are increasing exponentially in a very short period of time<sup>3</sup>. As we saw in the previous chapters collecting large amount of data is quick and easy. Technological advancements have enabled us to be able to think about utilising such large assemblages of data which would have been impossible even in the recent past. By providing unprecedented coverage, these large assemblages of data - 'Big data', provide us with insights which were not possible before. They often change our approach and methods employed in entire disciplines. For example, In computer science, fuelled by the explosion of collected user data, there is a paradigm shift in

<sup>1</sup> Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35 (2):137–144, 2015

<sup>2</sup> "Write programs that do one thing and do it well. Write programs to work together. Write programs to handle text streams, because that is a universal interface.", Doug McIlroy on UNIX philosophy.

<sup>3</sup> Rob Kitchin. Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1): 2053951714528481, 2014

Artificial Intelligence with the use of data mining, machine learning and deep learning. It is only time before this approach pervades social sciences research as well. In addition to the above advantages, Big data because of their nature also introduce numerous challenges in their collection, storage, analysis and visualisation. This is not including the enormous additional overhead and complexity introduced when we try to employ big data methods and tools. If we are not careful, using big data tools and methods for solving 'normal' problem can be counter productive where the advantages realised don't justified the overheads introduced. Hence it is important to understand the 'Big data' nature of the datasets we are dealing with at a granular level and choose the tools and methods without any presumptions.

The first and foremost challenge we face while discussing big data is its definition. It is hard to clearly and objectively define 'Big data' as it can vary widely based on the discipline and perspective. What may be 'big' in one discipline may not be in another. The nature of data can also be evaluated in various dimensions and can exhibit different properties in those dimensions. 'Big data' is generally defined within the context of disciplines, as data which cannot be managed with traditional methods and tools in those disciplines and requires substantial change in the approach of the practitioners. This definition is too subjective and falls short of giving us more understanding of 'Big data'. One of the most subscribed definition is to define the scale of the data in the dimension of volume - size of the data, velocity - speed of the data and variety - the complexity of the data<sup>4</sup>. This has also been extended to include more dimensions such as, veracity - the reliability or truthfulness of the data, visualisation - the complexity in visual interpretation and presentation of the data, and others such as visibility, validity, variability, volatility and value. There have also been other alternative dimensions proposed such as Cardinality, continuity and complexity<sup>5</sup>. However we can consider the core dimensions of data - volume, velocity, variety, veracity and visualisation for evaluating our datasets. Since not all data is 'Big' in all these dimensions, we need to evaluate the 'bigness' of the data in each dimension and consider the associated challenges and solutions.

The second set of challenges arise while we process the big data, its acquisition, storage, extraction, cleaning, annotation, integration, aggregation, modelling, analysis, visualisation and interpretation. Challenges in each one of these processing activity arises due to the data being big in one or more dimensions. The data being big in volume, velocity and variety poses challenges in data acquisition, aggregation, cleaning and analysis<sup>6</sup>. These challenges make traditional methods impractical and introduce the need for distributed, crowdsourced collection of data, heavily parallelised computing and application of functional programming concepts. The unstructured nature of the big data also introduces notable biases which mandate careful consideration, proper calibration and weighting during analysis so that we can understand and remove any uncertainties arising from them. The data being big in veracity dimension poses significant challenges in its analysis and modelling. Since simple methods such as linear regression fails in such scenarios, we require complex methods such as support vector machines, neural networks and hidden Markov models which compensate the lack of structure with the volume of data. With such computationally intensive meth-

<sup>4</sup> Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1, 2001

<sup>5</sup> Shan Suthaharan. Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review*, 41(4):70–73, 2014

<sup>6</sup> Songnian Li, Suzana Dragicevic, Francesc Antón Castro, Monika Sester, Stephan Winter, Arzu Coltekin, Christopher Pettit, Bin Jiang, James Haworth, Alfred Stein, et al. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS journal of Photogrammetry and Remote Sensing*, 115:119–133, 2016

ods, heavily parallelised high performance computing techniques such as GPU processing become indispensable. We also face significant challenge in visualising such complex features and methods which not only supports critical decision making but also is indispensable in exploratory analysis. The volume and velocity of big data makes them hard to visually simplify and digest. They are especially complex to interpret in the time dimension unless presented in small parts. Geographic information systems do a good job in visualising complex geographic data but struggle to maintain legibility and meaning when dealing with the temporal dimension. The visualisations of big data need to be highly processed, simplified and interactive to present meaning to the viewer. They have to balance between functionality, aesthetics and performance. Finally, because of the variety, big data creates need for consistent, well engineered standards so that multiple approaches and tools can be employed in tandem.

Apart from these processing challenges, we also have management challenges associated with big data such as privacy and security, data governance and ownership, data and information sharing, and cost<sup>7</sup>. Since these big datasets are usually comprehensive, securing them and protecting the privacy of the users becomes a central consideration in any project dealing with them. In many cases, though the data collected itself may not contain personal information but at these scales, in conjunction with other datasets, it can be used to infer them. The overall approach, methods, tools should comply with relevant legislation such as GDPR as well as the research ethics of all the stakeholders. This is especially challenging since these large unstructured datasets exhibit ambiguity of their ownership as well which calls for a clear, transparent and secure way to share them with other stakeholders along with publications of results in a timely, accessible manner. The associated project management and tracking tools need to be capable of handling these data ownership and sharing concerns as well.

Finally, the biggest challenge we face with big data is the cost in terms of money, resources and time. Though most of the big data tools are developed openly and distributed freely there can be lot of incidental, non-direct costs associated with collecting, processing and managing data with them. For example, there are the operational costs collecting data at such scale, network costs moving them, server costs storing and processing them, cost of procuring and supporting specialised tools and the human resource cost in hiring and training people who are capable for dealing with them. Though there are economies of scale at larger scales, the overall resources required to manage big data effectively can be several folds of what is needed for a traditional dataset. This makes it important to look at the data in our hands closely and carefully so that we can make informed decisions on how 'big' it is and choose the methods which are the most suited for such dataset.

#### *4.1.2 How big are the Wi-Fi probe request datasets?*

In this section we take a detailed look at the three sets of Wi-Fi probe requests collected as described in chapter on data collection using the 5Vs big data framework. Our aim is to understand the nature of the data in each dimension and thus evaluate the challenges we face in that specific dimension leading to

<sup>7</sup> HV Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94, 2014

a bespoke solution. We look at each set of data in each dimension and try to answer the following questions,

1. How can this dimension be measured objectively?
2. How big is the data in terms of the defined measurement?
3. How does it data compare with datasets in other disciplines?
4. How can we describe the size of the data?

We then combine these isolated evaluations to form a combined description of the datasets. This is then used as the basis for developing a list of requirements for designing the data processing and management toolkit.

#### *Volume*

Probe requests data, being dynamic and continuous, cannot be quantified as an absolute static number in terms of volume. Hence we use a long term measurement - yearly rate, for each location instead. On shorter datasets such as the pilot study, we estimate the yearly volume linearly from the available data. We standardise this measure as the amount of disk space needed to store the collected data when encoded in text form. It is important to note that this can be reduced many folds by using compression or binary formats but we chose text since it is the de-facto standard for exchanging data.

Table 4.1: Comparison of volume or size of the datasets of WiFi probe requests.

Study	Maximum*	Minimum*	Average*	Total**
Pilot Study	134	3	54	48.3
Main Study	6.1	2.4	4.42	4.1
Smart Street Sensor	5.4	0.001	0.8	0.8

\* Measured/ Estimated for each location in gigabytes per year. \*\* Measured/ Estimated for 920 locations in terabytes per year

We can see that there is a lot of variability in the volume of probe requests generated at a given location. This mostly depends on how many mobile devices are present around the location. We observe that when we collect most of the information present in the probe requests in a busy area such as Oxford street in the Pilot studies, we generate around 50 terabytes of data in a year. But in a more real world setting such as the Smart Street Sensor project where the sensors fail at times and the amount of data collected is optimised, the volume is around a 1 gigabyte. The total volume of data we deal with in the case of a national scale project with around 920 sensors running for around 4 years is around 2 terabytes. A comparison of this to datasets from other disciplines is shown in Figure 4.1. It is key to note that the y-axis is scaled exponentially.

We can see that the probe requests data is not truly 'Big data' as experienced in other fields. It is only when we reach a complete coverage, i.e, putting a sensor at each retail establishment in UK, our estimated data volume reaches around 250 petabytes which is comparable to scales experienced in other fields such as particle physics and world wide social networks. At the same time, the scale of probe request data is not small either. The volume of 2 terabytes is more than the memory available in any desktop systems and is more than any of them can process in a timely manner. Summarising from the above, we can confidently say that the probe request datasets are 'Medium Data' - especially the dataset collected by the smart street sensor project. Though it

has potential to scale into a truly big dataset, for the purposes of this research we can classify it as 'Medium data' in the volume dimension.

### Velocity

Velocity is the rate at which the data is collected over time. It is significant when evaluating big data since some data which may not scale in terms of absolute volume but the speed at which they are received makes them challenging to deal with. A perfect example is the comparison between data generated by the Large Hadron Collider project by European Council for Nuclear Research and a world wide social network such as Facebook. Though their total volumes are comparable at 200 petabytes, the data from LHC is generated in concentrated experiments at a rate of 3 petabytes in 5 seconds while Facebook generates the same about in about a day or two. Since the size of an individual Wi-Fi probe request doesn't vary widely, we define the velocity of this dataset as the number of requests received at a given location at a given location within a given time interval. Though the precision of the time measured during data collection is in microseconds, the practical data transfer resolution in all the datasets is around 5 minutes. Hence we measure velocity of our datasets in terms of number of requests every 5 minutes. Table 4.2 compares the datasets we collected on Wi-Fi probe requests in terms of their volume.

Study	Maximum*	Minimum*	Average*	Total**
Pilot Study	8577	188	3469	3.20
Main Study	1362	534	782	0.72
Smart Street Sensor	5024	6	408	0.27

We observe that locations can receive up to 8500 requests in 5 minutes or can get no request at all depending on the time and how busy it is. But we can see that on average a national scale project with around 900 locations generates around a million requests every 15 mins. Compared to the LHC's 180 billion records or Google's 190 million searches per 5 minutes this seems to be not high speed data. However, this is much faster compared to traditional data sources such as census or geographical surveys which are updated anywhere between 6 months to 10 years.

To summarise, in terms of velocity, the Wi-Fi probes data can be described as 'Medium' at best. The methods dealing with the data should be time sensitive and be able to deal with a continuous stream of data but at the same time need not be real time or need sub-second latency. Since the Wi-Fi probe requests don't have actual location information the mobile devices, it does not have the similar value in real-time analytics as shown in comparable location or movement based datasets.

### Variety

Variety is defined by the amount of variance in the type and characteristics of the data. Since variety is hard to quantify and compare across disciplines we evaluate the dataset subjectively for the variety present in it. The data transmitted in a Wi-Fi probe request is defined by the 802.11 Wi-Fi specification

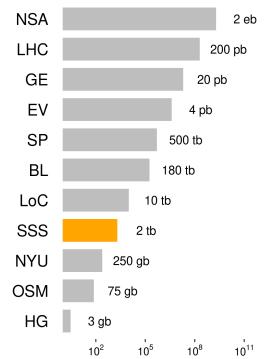


Figure 4.1: Comparison of volumes of datasets across various disciplines.

NSA - National Security Agency,  
LHC - Large Hadron Collider,  
GE - Google Earth, EV - Event Horizon project, SP - Spotify music,  
BL - British Library data store, LoC - Library of Congress,  
SSS - Smart Street Sensor, NYU - New York city Uber trips 2009-15,  
OSM - Open Street Map and HG - Human Genome Project

Table 4.2: Comparison of velocity or speed of the datasets of Wi-Fi probe requests.

\* Measured/ Estimated for each location in number of requests per 5 minutes. \*\* Measured/ Estimated for 920 locations in Millions of requests per 5 minutes

<sup>8</sup> IEEE. IEEE standard for information technology-telecommunications and information exchange between systems local and metropolitan area networks-specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. *IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)*, pages 1–3534, Dec 2016

<sup>9</sup> Mathy Vanhoef, Célestin Matte, Mathieu Cunche, Leonardo S Cardoso, Frank Piessens, and Piessens Frank. Why mac address randomization is not enough: An analysis of wi-fi network discovery mechanisms. In *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*, pages 413–424. ACM, 2016. ISBN 1450342337. DOI: 10.1145/2897845.2897883

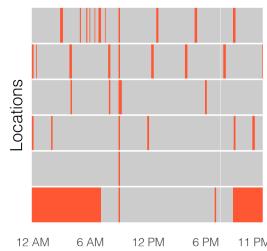


Figure 4.2: Missing data from five locations at Tottenham Court Road, London on 15 January 2018 demonstrating the veracity of the data.

<sup>8</sup> and every probe request has to have a set of mandatory fields for Wi-Fi to work. This set of fields is the same everywhere across the world and the specification, especially the probe request part, has remained stable over years. Though there is some variability allowed within the specification, being part of a global standard, the data collected is heavily structured in general.

The first set of variety present in the Wi-Fi probes data set arises from the ‘information elements’ part of the probe request. The structure of a probe request is discussed in detail in the data collection chapter and is summarised in Figure ?. Essentially the information about the capabilities and type of the mobile device is encoded in the information elements part of the probe request and this information is optional and is implemented at the discretion of the manufacturers. As this information elements are demonstrated to be useful in successfully fingerprinting the mobile devices <sup>9</sup>, mobile devices increasingly don’t include any information in them. Emergence of manufacturers with large market share and narrow set of device models such as Apple and Samsung also reduce further variability in them. The second set of variety in the dataset arises from the rate at which these probe requests are generated by the mobile devices. Unlike devices which generate data on events or at regular intervals, mobile phones generate probe requests at a rate based on various factors. Though this leads to some challenges in counting footfall from these probe requests the variability exhibited here is neither so large nor so complex that traditional methods could not deal with them.

Comparing with some of the big data encountered in unstructured data collected over web such as social networks or other sensor based methods, the variability here can be considered trivial. Further when we convert these probe requests in to footfall counts, the variety in the dataset drops almost to zero as it becomes just an ordinal data point varying in geography and time. Summarising the above, we can confidently say that the Wi-Fi probe request data does not exhibit any ‘big data’ properties in the variety dimension.

### Veracity

Veracity is defined as the amount of abnormality present in the data in the form of inaccuracies, biases and noise. Similar to variety, veracity is hard to quantify hence required a subjective evaluation. Being sensor collected data, veracity is the dimension where the data exhibits most ‘big data’ properties.

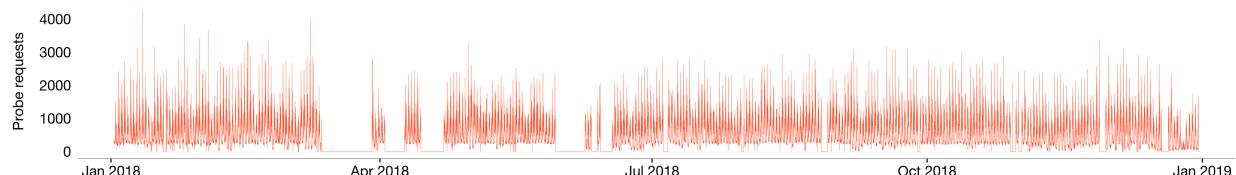
First set of veracity in the dataset arise from the fact that it is collected through sensors located in multiple locations which communicate to the central server using 3G mobile data connectivity. We know from experience that the sensors are unreliable and fail to send back data regularly due to various reasons. More over the sensors are installed and uninstalled regularly as partners join and leave the project. This results in a data stream which is often erratic and incomplete with large gaps in them. In addition to this the sensors need to be rebooted regularly due to issues or updates leading to small gaps as well. Since the sensors are part of retail establishments they can be switched on and off regularly in some of them as well. Figure 4.2 demonstrates the veracity of the data in terms of missing data for a sample of locations in London. All the above pose immense challenges when we attempt to aggregate the data where we have to estimate and fill these gaps.

There is also a lot of variability in the physical location of the sensors and the area of measurement. The sensors may report higher or lower count due to their configuration and the context of their location as discussed in chapters pertaining to data cleaning. This leads to a situation where the accuracy of the data collection varying quite widely across location and times<sup>10</sup>. It is often not clear if the change in the data is due to actual changes at the location or just the change in the configuration of the device. For example, Opening of a mobile shop next door to the sensor can increase the estimated footfall without any change in actual footfall at the location.

Finally we also have to work within the changing mobile landscape. Though the Wi-Fi probe requests are standardised by IEEE, the mobile manufacturers have started adopting obfuscation techniques to protect the privacy of the users. This started with randomisation of MAC addresses, removal of information elements and generally getting more sophisticated with new versions of operating system. There is also bias in terms of operating system adoption and change in market share between manufacturers. There is no inherent structure or information on what is changed and how often these changes occur which leads to questions on the continuity of the data over long periods of time.

Summarising from the above, we can confidently conclude that Wi-Fi probe requests dataset shows 'Big data' characteristics in terms of its veracity and requires appropriate tools and methods when aggregating, analysing and modelling it.

### *Visualisation*



Visualisation is closely related to volume, velocity and variety of the data. The Wi-Fi data due to its non-trivial volume and velocity, exhibits similar characteristics and challenges in terms of visualisation. Since there is not much variety in the dataset, when we process the raw data into footfall counts we are left with just the time, location and footfall count for each data point. Out of these, location and footfall counts are easy to visualise but time exhibits big data properties. This is primarily due to its granularity at 5 minute intervals and longitudinal nature of the data collection. The major challenge with Wi-Fi data is to simplify and visualise them in a legible way while showing change in term of time. The veracity of the data presents challenges in simplifying them and the volume poses challenges in maintaining legibility. We also have to take the 'near real time' aspect of the data into consideration while visualising them. There is a clear need for always on, interactive, real time dashboards with geographic capabilities in addition to the capabilities of traditional desktop GIS. There is also need for multiple linked dynamic visualisation platform

<sup>10</sup> Karlo Lugomer, Balamurugan Soundararaj, Roberto Murcio, James Cheshire, and Paul Longley. Understanding sources of measurement error in the wi-fi sensor data in the smart city. In *Proceedings of GISRUK 2017*. GIS Research UK (GISRUK), 2017

Figure 4.3: Number of probe requests collected for every five minute interval at Tottenham Court Road, London on the year 2018 showing the visual complexity of data in the time dimension.

for separating the scope of the visualisation into manageable units. Figure 4.3 demonstrates the illegibility of simple visualisations of the data due to granularity, variability and veracity. We can safely say that the Wi-Fi probe requests dataset is at best 'Medium' in the visualisation dimension.

Summarising the above discussion, we can conclude that the datasets collected from Wi-Fi probe requests are at best of 'medium'. They show the most big data characteristics in terms of their veracity. In rest of the dimensions the datasets are not truly big data and we need to look at tools and methods appropriate to their size. The toolkit we devise need to be able to deal with their mid-size volume, velocity and visualisation dimensions and at the same time need to be able to deal with the large amount of veracity of in them. Figure 4.4 illustrates the summary our discussion. This leads us to devise a 'medium data toolkit' which can be used without incurring the extra cost and complexity introduced by big data tools while be able to handle the data at hand.

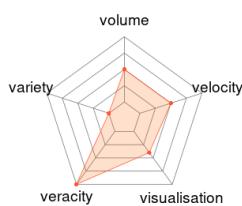


Figure 4.4: Big data characteristics of the Wi-Fi probe request datasets in their corresponding dimensions

#### 4.1.3 A Survey of Methods and Tools

Having classified the Wi-Fi probes dataset as a 'Medium' sized data, in this section, we survey the tools and methods available at various stages of the data processing and management process - data collection, storage and retrieval, processing and analysis, visualisation. We first survey the tools available in each stage and specifically look at their suitability for Wi-Fi probe request datasets in terms of the following characteristics,

- *Performance* - How much data can be processed in a given time?
- *Flexibility* - How easy it is to change the scale and scope?
- *Complexity* - How many components or parts are involved?
- *Cost* - How much money or infrastructure do they require?

We then discuss the principles of UNIX philosophy and how it helps in solving similar sized problems in computer science. Finally we pick and connect the tools to devise our toolkit which is best suited for our Wi-Fi probe request dataset.

#### *Collection*

We discussed various technologies used in collecting passive data on ambient population and pedestrian movement in the literature search. In this section we look at tools and methods used to collect Wi-Fi based data passively. The primary considerations for evaluating data collection strategy are the scale of the infrastructure, expertise and effort required to implement it and cost involved.

There have been numerous sensors, tools and associated software platforms made available for data collection under the umbrella of 'internet of things'. We start by looking at different approaches in the Wi-Fi data collection tools and try to reason the most appropriate solution for our research. On one end there are low level and low cost bespoke solutions which require lot of effort to implement and maintain. On the other end there are turn key solutions which doesn't require lesser effort but costs considerably more. The key is finding a

balance between both while satisfying the requirements of the project. Since the Wi-Fi data is medium sized in terms of volume and velocity, we can deal with solutions with less than optimal scalability but since the data is 'big' in terms of veracity the toolkit has to give us most flexibility. Essentially, we are looking for a data collection methodology which prioritises flexibility and cost while performing moderately in terms of scalability and complexity as illustrated in Figure 4.5.

Type of solution	Examples
Bespoke	Micro-controllers with Wi-Fi modules e.g. Audrino + ESP8266
Turn-key	End to end commercial services e.g. Blix, Euclid, Pygmalios etc.
Ideal	General purpose hardware e.g. Raspberry Pi, Repurposed mobile devices - Tablets, Phones etc.

Table 4.3: Examples of different types of Wi-Fi based data collection solutions.

In terms of hardware, an example of a highly customised solution would be a micro-controller, such as Audrino, coupled with dedicated Wi-Fi module and programmed with custom software to collect the exact data needed. Designing and implementing of such system is time consuming, cumbersome and usually involves significant cost but it can also be highly flexible, efficient and cheap to deploy. On the other end of this spectrum, we have end-to-end solutions such as Blix, Walkbase, Ecuclid, Retail next, pygmalios etc. where the data is collected through multiple sensors and sources and syndicated into a clean footfall information by a third party service provider. These platforms for footfall data collection and analysis have the advantage of being quick and easy to develop and deploy while they can also be highly inflexible for changes and turn out to be costly when scaled up. A middle ground here is to use a general purpose hardware such as single board computers or repurposed mobile devices, augment them with additional hardware modules and use general purpose scripting languages to write software for them. This way we avoid low level hardware or software design and implementation while maintaining good amount of flexibility. Table 4.3 shows some examples of such systems while highlighting an ideal system.

The Smart Street Sensor project uses its own proprietary sensor system designed and instrumented by the data partner. The design and implementation decisions were made with the commercial application in mind and is not entirely relevant to our discussion in the context of our research. For the research conducted with the data, it is necessary to understand the data collection process and make sure it aligns and integrates with the rest of the toolkit. As discussed in the data collection chapter, the methodology used in the smart street sensor project satisfies our requirements. The toolkit we designed to collect other datasets are in-part inspired by this methodology or a modified version to include more flexibility. The toolkit consists of Raspberry Pi, Linux, tcpdump or tshark <sup>11</sup> and nodejs. Raspberry Pi and the Linux OS provides a general purpose base system and hence the flexibility. On top of this we built our data collection system by assembling open source and free network analysis tools such as tcpdump and tshark along with other tools providing functions such as scheduling, personal data obfuscation and data

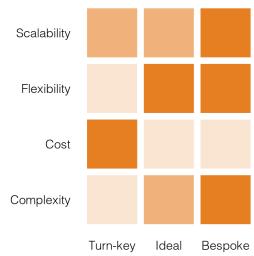


Figure 4.5: Characteristics of types of Wi-Fi data collection tools at each end of the spectrum compared to an ideal candidate  
(Darker colors show higher score)

<sup>11</sup> Gerald Combs and Contributors. Wireshark - network protocol analyzer. <https://www.wireshark.org/about.html>, 2018

transmission with scripting languages like nodejs and bash.

### *Storage*

Data storage technology is one of the most diverse landscape in terms of both methods and tools available. It has been constantly in research and development since the beginning of computing and is one of the fastest changing landscapes with the advent of big data paradigm. A comprehensive review of storage solution warrants a chapter in itself so we restrict our survey to an outline of most significant approaches and corresponding systems and tools.

At one end of the spectrum is one of the most underappreciated for of data storage - File systems. Though they seem like a low level interface for storing data, file systems have their advantages as well. When the data is not complex or inter-related, flat text files in file systems could be the fastest way to store, search and retrieve data. Since operating systems are usually optimised to manage storage media through file systems, they involve no additional overhead and are extremely reliable. The hierarchical file systems use in most of the operating systems act as an index with hierarchical data. The major disadvantage of file systems is that they are not useful for managing data with any kind of complexity. This is the primary reason why database management systems are developed on top of file systems.

Database systems can be broadly divided into relational and document based. The relational databases are optimised to deal with relational data and usually enforce strict structure for the data In general they can handle large number of rows and are designed to scale vertically. Most relational database systems try to guarantee ACID <sup>12</sup> compliance and hence used in critical systems such as financial operations, sales etc <sup>13</sup>. The document based databases are optimised to deal with unstructured data and can doesn't need a strictly defined schema. In general they can handle large number of columns and are designed to be distributed and scaled horizontally. Being distributed, most document based databases try to pick a focus and compromise on others as specified in CAP theorem <sup>14</sup>. There are numerous databases systems which prioritise different things and the right solution depends on the properties of the data and the requirements of the project.

Since the publication of the paper on 'Google file system' by google <sup>15</sup>. There have been significant effort in designing and building 'big data' file storage systems which can store large data in the range of petabytes. These systems are designed to be distributed and optimised for high throughput for queries on them. Hadoop Distributed File System (HDFS) is one such file system which is also the most widely adopted. There are numerous cloud based, third-party solutions built with these file systems making them easy to use. There are also numerous tools, libraries and frameworks which emulate the features of database systems on these distributed file systems making them easier to use further. The primary advantage of these systems is the sheer scalability they provide when it comes to data volume. The primary disadvantage is the associated overheads in terms of cost and time incurred in learning, designing and implementing them. Unless the project is sufficiently large, the advantages gained usually do not justify the overheads introduced.

<sup>12</sup> Atomicity, Consistency, Isolation and Durability are properties which make sure that the data in the database is valid even during failures.

<sup>13</sup> Theo Haerder and Andreas Reuter. Principles of transaction-oriented database recovery. *ACM Comput. Surv.*, 15(4):287–317, December 1983. ISSN 0360-0300. DOI: 10.1145/289.291. URL <http://doi.acm.org/10.1145/289.291>

<sup>14</sup> Brewer's theorem or CAP theorem states that it is impossible to simultaneously guarantee consistency, availability and partition tolerance in a distributed data store.

<sup>15</sup> Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *Proceedings of the 19th ACM Symposium on Operating Systems Principles*, pages 20–43, Bolton Landing, NY, 2003

Table 4.4 summarises the above discussion along with relevant examples.

Approach	Data size	Examples	Comments
File system	< 10 TB	ext, ntfs, zfs, btrfs	Simple and efficient. Best for hierarchical data. Cannot handle complex connected data.
Relational DB	< 5 TB	MySQL, PostgreSQL	Handles structured and relational data. Optimised for large amount of rows and tries to guarantee validity.
Document DB	< 10 TB	MongoDB, Cassandra	Handles unstructured data. Optimised for large number of fields and distribution to multiple clusters. Tries to focus on any two guarantees of the Brewer's theorem.
Distributed FS	> 10 TB	HDFS, GFS	Ceph, Optimised for really large datasets which need to be distributed over multiple nodes.
Cloud Storage	> 10 TB	AWS, SWIFT	Azure, Implements distributed file systems on the cloud. Has more reliability and scalability than local implementations.
Data Warehouse	> 10 TB	Hive, Impala, Presto	Hbase, Interfaces built on top of distributed file systems to emulate capabilities of relational databases on them.

Table 4.4: Various data storage approaches and their characteristics.

We saw that the Wi-Fi probe request datasets are 'Medium' sized hence we can safely eliminate distributed file systems for storing them. Though the smart street Sensor project uses Azure Blob Storage, when the data is downloaded to the local servers at the university, we can just store them in the file system because of their size (2TB approx.) and the hierarchical nature. The folder structure of - year/month/day/location/interval/ with individual text file, enable us to query the data for any given location at any interval nearly instantaneously without any further database operations. When this raw data is processed into 5 minute counts, we require more relational queries. For this purpose a relational database is sufficient as volume is quite small (25GB approx.). We chose PostgreSQL because of the PostGIS extension which gives us flexibility in handling geographic data.

### Processing

We saw that the data is medium in terms of volume and velocity and shows big data properties in terms of veracity. Hence we require tools which are capable of dealing with the veracity of the data while being able to manage the volume and velocity. The traditional approach to deal with such dataset is to load it into a general purpose analysis tool such as R or a GIS packages and process it. The size of the dataset and the lack of meaningful complexity of geography element in the data eliminates the use of GIS packages. Scripting languages such as R and Python can deal with the dataset and its requirements but the time taken to do so increases exponentially with the size of the data as the size of objects in memory increases. Figure 4.6 illustrates the increase in processing time with respect to number of location for a simple exercise where a day's worth of raw data is parsed and aggregated into number of probe

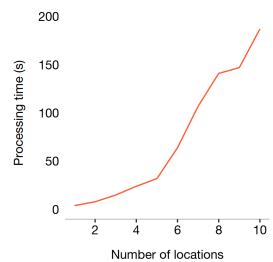


Figure 4.6: Exponential increase in the processing time when using traditional methods.

*The processing involves parsing JSON data received for a single day at each location and aggregating them as number of probes requests received in every five minute intervals.*

requests per 5 minute intervals (The code used to produce these benchmarks are detailed at Section 7.1). This becomes prohibitively expensive as the number of locations and complexity of the processing increases. Though this can be improved with more efficient coding practices, the margin of improvement is quite limited hence creating the need for better techniques. It is important to note that data processing is done in two stages - the first stage where the raw Wi-Fi probe requests are filtered, cleaned and aggregated into footfall counts and second stage where the footfall counts are in turn analysed to produce reports and visualisations. The traditional methods are sufficient for the second stage of the processing and the first stage is the one which requires a better solution.

On the other end we have big data analysis tools which are built for dealing with extremely large amount of data. Since the publication of the paper on MapReduce, there have been immense developments in the Big data analysis landscape. There numerous distributed programming tools to use the data stored within a distributed storage system each focussing on specific type of data and analysis. A concise, non-comprehensive list of types of data or specialities and corresponding big data tools is shown in Table 4.5.

Table 4.5: Various types of big data processing tools and corresponding examples.

Tools	Speciality
General purpose	MapReduce, Spark
Real-time streams	Flink, Pulsar
Events or messages data	Storm, Kafka, Flume
Networked or graph data	Tinkerpop, Corona
Scheduling	Oozie, Falcon, Azkaban
Turn-key platforms	SpringXD, Cask Data

We can rule out the necessity of the above big data tools since our dataset is neither big enough nor fast enough. The dataset does not have any specialised structure such as graph or network but just a very minimal component of geography to it. Using any of specialised big data tools is just going to introduce immense overheads without any added benefits. We need something in-between the above two approaches where we is sufficiently fast and flexible for our datasets.

<sup>16</sup> Command-line tools can be 235x faster than your hadoop cluster, Jan 2014. URL <https://bit.ly/2s2XZYI>

<sup>17</sup> Malcolm D McIlroy, Elliot N Pinson, and Berkley A Tague. Unix time-sharing system: Foreword. *Bell System Technical Journal*, 57(6):1899–1904, 1978

This is where we come across the possibility of using standard Unix tools along with connecting them to create a processing pipeline. In some cases, a data processing pipeline made using command line Unix tools have been demonstrated to be 230 times faster than using big data toolkits <sup>16</sup>. The command line tools were developed as parts of Unix operating system for processing text. They are developed in line with the Unix philosophy which focuses on modular and minimal software development. The core tenants of the Unix philosophy has been summarised by Doug McIlroy as below,<sup>17</sup>,

1. Make each program do one thing well. To do a new job, build afresh rather than complicate old programs by adding new "features".
2. Expect the output of every program to become the input to another, as yet unknown, program. Don't clutter output with extraneous information. Avoid stringently columnar or binary input formats. Don't insist on interactive input.
3. Design and build software, even operating systems, to be tried

early, ideally within weeks. Don't hesitate to throw away the clumsy parts and rebuild them.

4. Use tools in preference to unskilled help to lighten a programming task, even if you have to detour to build the tools and expect to throw some of them out after you've finished using them.

These principles along with the ‘pipe’ operator gives us necessary tools to build more complex tools. We can replace most of the libraries we used in the R implementation of our processing with a corresponding command line tools and connect them together with a text interface to achieve similar pipeline. The first advantage of such design is that it is much more efficient than a monolith design. These tools being actively developed for since their invention are compiled as native binaries and are usually extremely optimised resulting in a much faster pipeline. Because of the design of the pipe operator, the individual parts of the pipeline are executed in parallel as chunks of data are passed through them thus avoiding the need to load entire datasets into memory which results in an exponential increase processing time with the size of the data. Being modular, we can even introduce process level parallelism to parts of the pipeline without any major change in the overall design. Finally the modular structure also gives us the advantage of using the best tool for any part of the pipeline.

Tools	R Library	Unix tool(s)
Move data to and from Azure blob storage, SQL server and Postgres	AzureR, odbc , RPostgreSQL	azcopy, mssql, psql
Convert data from JSON format to CSV	jsonlite	jq
Encrypt raw data for secure storage	Rcrypt	gnupg
Anonymise personal data into cryptographic hash	digest	openssl
Transform and manipulate tabular data	dplyr	find, cat, cut, grep, sed, awk, sort, uniq, column, paste, join
Impute missing value using time series analysis	imputeTS	Rscript
Visualise the results into maps and charts	ggplot2	Rscript
Create and manipulate geographic data	sf, rgdal	postgis, gdal

All of the gives us an extremely minimal and efficient toolkit to process the raw Wi-Fi probes data into counts in a scalable way. Figure 4.7 compares the processing times of such Unix toolkit with the traditional R based toolkit as the data size increases. We can see that Unix toolkit performs extremely well and the performance gains are significant as the size of the data increases. For example, to process data for 25 locations, R based toolkit takes around 20 minutes while the Unix toolkit gets it done in 20 seconds (The code used to produce these benchmarks are detailed at Section 7.1). Table 4.6 shows the activities in our pipeline and corresponding libraries in the traditional R workflow along with the equivalent Unix tools. It is important to note

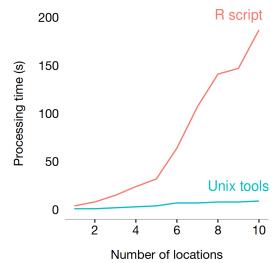


Figure 4.7: The increase in processing time with the Unix pipeline is linear thus improves the scalability compared to R based processing

Table 4.6: Tasks in the processing pipeline, corresponding R libraries and equivalent Unix tools

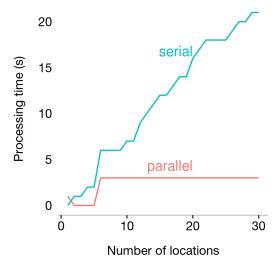


Figure 4.8: The scalability of the processing pipeline could be further improved with parallelising it.

that tools for doing specialised actions such as statistical analysis, machine learning and time-series analysis are built on top of scripting languages such as R and Python.

<sup>18</sup> Ole Tange. *GNU Parallel 2018*. Ole Tange, March 2018. ISBN 9781387509881. doi: 10.5281/zenodo.1146014. URL <https://doi.org/10.5281/zenodo.1146014>

<sup>19</sup> Stephen R. Walli. The posix family of standards. *StandardView*, 3(1):11–17, March 1995. ISSN 1067-9936. doi: 10.1145/210308.210315. URL <http://doi.acm.org/10.1145/210308.210315>

<sup>20</sup> Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011. URL <http://vis.stanford.edu/papers/d3>

These can be embedded into our Unix pipeline as scripts running in corresponding front-ends such as Rscript or python. This toolkit can be further accelerated by parallelising parts of the pipeline using gnu-parallel<sup>18</sup>. For example, the previous example pipeline can be parallelised by spawning a pipeline for each location this reduces the processing time for a set of 25 locations from 18 seconds to 3 seconds. This done by utilising every processor cores available in the CPU. Figure 4.8 compares the processing times of the Unix toolkit with a parallelised implementation (The code used to produce these benchmarks are detailed at Section 7.1). Finally all the Unix tools discussed in this toolkit are open source and free software which has almost no cost in terms of resources. Since these tools are part of the POSIX specification<sup>19</sup> for operating systems, the expertise in their design and use are transferable to and from other disciplines thus reducing researcher time learning and using these tools.

#### *Visualisation*

In the last section we saw that the visualisation dimension of the data shows some level of complexity. The primary source of this complexity arises from the longitudinal nature of the data and the noise due to granularity of the data. For the processed dataset, traditional visualisation and mapping libraries with R is sufficient while the visualisation of raw data across long time periods for either for exploratory analysis or for communication needs some form of interactivity or simplification to be able to legible. Data driven documents (D3)<sup>20</sup> and Dimensional charting (DC) provides us with both of these requirements. Both of these tools can accept text based input and can fit with other Unix tools discussed earlier. In case of binary file output such as images or documents, they could be directed to the file system and then read into other programs.

#### *4.1.4 The Bespoke ‘Medium data toolkit’*

In this chapter we saw how the advent of internet and internet enabled devices has lead to significant increase in the amount of data generated and collected across disciplines. This data deluge and improvements in the capabilities of computing hardware has fuelled an explosion of research and development in tools and methods to deal with these ‘Big data’. Though these big data tools promise huge improvements in processing capabilities, when used under wrong circumstances they can lead to unwanted overheads and costs. Thus we need a framework to examine and understand the scale of the data that is being used so that we use the right tools for the right purposes and the 5Vs of ‘Big data’ - Volume, Velocity, Veracity, Variety and Visualisation provides us with such frame work. Every dimension of big data poses unique set of challenges and we need make right decisions in choosing specialised tools and methods to overcome them.

We then closely examined the Wi-Fi probes data we collected with this framework and found that the data, though posed significant challenges with

traditional data processing techniques, do not exhibit 'big data' properties in all its dimensions. Only veracity of the data was found to have any meaningful big data properties, while volume and velocity was found to be 'medium' at best. The datasets lacked any variety and posed minimal challenge in the visualisation dimension because of its high temporal granularity. Thus we arrived at the requirements for a bespoke 'medium data toolkit' which is able to deal with these challenges.

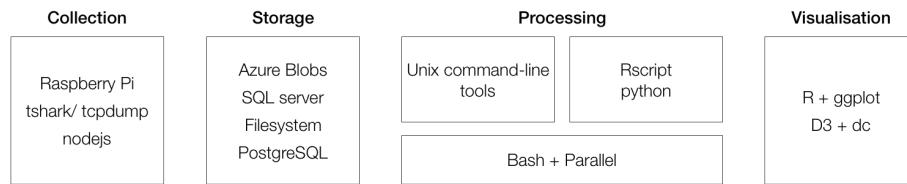


Figure 4.9: Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data

We undertook a brief survey of tools available for collecting, storing, processing and visualising the Wi-Fi probe request data and with the understanding of the data from the previous analysis chose the ones which are perfect for the datasets. For collecting Wi-Fi probes data in a scalable way, we chose a general purpose single board computers such as Raspberry-Pi along with open source tools such as tcpdump and tshark in a Linux environment. For data storage we narrowed in on using just the file system for the raw data and relational database management system for the processed counts. To process the raw data we chose to devise a processing pipeline using an assortment of standard Unix command line tools linked together using a shell scripting language and parallelised at the process level with gnu-parallel. We also demonstrated that this processing pipeline can be 400 times faster than the using a monolithic pipeline even with a small sample of locations. For visualisation we chose D3 and DC as the solution for communicating time information legibly. Finally we arrive a 'medium data toolkit', illustrated in Figure 4.9, which is best suited for the Wi-Fi probes dataset which we employ to process and examine the data further.



## 4.2 Device fingerprinting

In the past decade, Wi-Fi has emerged as one of the most commonly used technologies in providing high speed internet access to mobile devices such as smartphones, tablets and laptops in public and private spaces [Torrens, 2008]. This has resulted in multiple Wi-Fi networks being available at almost every location in dense urban environments. Traversing through this overlapping mesh of Wi-Fi networks, modern mobile devices with Wi-Fi network interfaces regularly broadcast a special type of signal known as ‘Probe Requests’ in order to discover the Wi-Fi networks available to them. This helps these devices to connect and switch between the Wi-Fi networks seamlessly.

Probe requests are low level signals standardised by IEEE 802. 1 specification [IEEE, 2016] for service discovery, and are implemented in any Wi-Fi capable device irrespective of the manufacturer or the model. This ubiquity and standardisation makes them an excellent source of open, passive, continuous, and wireless data generated by Wi-Fi capable devices present at any given time and location. Considering the unprecedented levels of mobile device ownership in recent years, we can, in turn use this data to understand the population distribution in highly dynamic urban environments with high spatial and temporal granularity [Freudiger, 2015, Kontokosta and Johnson, 2017]. While a Wi-Fi based method to collect data offers us various advantages such as, easy scalability and efficiency in terms of cost and time, it also introduces few systematic biases and uncertainties in the collected data along with the serious risk of infringing on the privacy of the mobile users. In this section, using the set of probe requests and manual counts collected at various high street locations across London, we demonstrate that pedestrian footfall at these locations can be estimated with considerable precision and accuracy while protecting the privacy of the pedestrians.

Unlike GPS, the location of the Wi-Fi enabled mobile device cannot be directly inferred from Wi-Fi, however there are reliable methods to triangulate the location of mobile devices from the locations of known access points (AP) and the signal strength reported by them [He et al., 2003, Moore et al., 2004, LaMarca et al., 2005]. This can overcome the usual shortcoming of GPS, which struggles for precision and accuracy in indoor and densely built environments [Zarimpas et al., 2006, Kawaguchi, 2009, Xi et al., 2010]. Utilising this, we can easily and quickly estimate trajectories of the mobile devices [Musa and Eriksson, 2012] which can be used similarly to the GPS trajectories to understand individual travel patterns [Rekimoto et al., 2007, Sapiezynski et al., 2015], crowd behaviour [Abedi et al., 2013, Mowafi et al., 2013], vehicular [Lu et al., 2010] and pedestrian movement [Xu et al., 2013, Fukuzaki et al., 2014, Wang et al., 2016]. Such data can also be used in transportation planning and management to estimate travel time [Musa and Eriksson, 2011] and real time traffic monitoring [Abbott-Jard et al., 2013b]. Using techniques demonstrated by [Franklin et al., 2006] and [Pang et al., 2007b], along with information present in the probe requests, one can even model interactions between the users [Cheng et al., 2012, Barbera et al., 2013, Cunche et al., 2014] such as predicting which of them are most likely to meet again [Cunche et al., 2012]. Using the semantic information present in these probe requests it even is possible to understand the nature of population at a large scale [Di Luzio

et al., 2016].

Although extensive research has been carried out on this subject with feasible and favorable results, in recent years, one of the major challenges faced in such attempts has been the increasing attempt by mobile phone manufacturers to protect their users' privacy by anonymising the globally identifiable portion of the probe requests [Greenstein et al., 2008]. Various methods have been devised to overcome this anonymisation process such as estimating the device model information from a known dataset of manufacturers and device behaviours [Martin et al., 2016]; Scrambler attack using a small part of the physical layer specification for Wi-Fi [Vo-Huu et al., 2016, Bloessl et al., 2015]; and timing attack where the packet sequence information along with information elements present in the probe request frame is used [Matte et al., 2016, Cheng and Wang, 2016]. A combination of these methodologies has been proven to produce de-anonymised globally unique device information [Vanhoeft et al., 2016, Martin et al., 2017]. These approaches usually result in serious risk of breach of privacy of the users of the mobile devices by revealing their identifiable personal information.

There is a clear gap in the research for exploring methodologies for estimating the number of unique mobile devices from a set of anonymised probe requests, without the need to reveal their original device information. Such a technique has various applications such as uncovering the urban wireless landscape [Rose and Welsh, 2010], revealing human activity at large scales [Qin et al., 2013], estimating pedestrian numbers in crowds [Schauer et al., 2014, Fukuzaki et al., 2015], and even counting people in hyper local scales such as queues [Wang et al., 2013]. With enough infrastructure to collect such information we can even aim to generate a real-time census of the city [Kontokosta and Johnson, 2017]. With this background, we set out to devise and implement a methodology to reliably estimate human activity such as pedestrian footfall from Wi-Fi probe requests without risking a breach of privacy of the users involved.

### *Methodology*

The primary aim of this research was to enable us to collect a series of probe requests and process them into a usable pedestrian footfall count. We did this by using a Wi-Fi receiver to collect probe requests broadcast by mobile devices, filtering out the background noise, and aggregating them based on the device that generated them. In this section, we examine the characteristics of probe requests in detail, devise a methodology to collect these probe requests in public areas, examine the systemic biases and uncertainties in the data collection method, and devise data processing methods to overcome these challenges. Finally, we compare the processed footfall counts to the ground truth recorded by primary surveys.

Probe requests are a special type of management packet broadcast by Wi-Fi enabled devices as part of their various functions such as scanning for available APs and quick geolocation by triangulation based known APs, etc. These are broadcast by all Wi-Fi enabled devices regardless of the manufacturer, type or model of the devices, although there is some variation in the frequency and the content of the information transmitted through them. In some cases, such

as Android devices, these are broadcast even when the Wi-Fi functionality has been turned off by the user so that the device can immediately connect to networks when the functionality is switched back on. Since some devices even use the probe requests as a less accurate form of localisation, they continuously send probe requests when Wi-Fi has been switched off. Thus, these signals can be used to reliably identify the presence of Wi-Fi enabled mobile devices. Being a first step of connection initiated by the mobile device, these packets have information regarding the characteristics of the mobile device itself. Some of the key information we can infer from these requests are,

1. **Media Access Control (MAC) address** which is a name identifying the wireless hardware of the mobile device,
2. **Sequence number** of the request for the mobile device to keep track of the responses,
3. **Time stamp** at which the request was received by the AP,
4. Total **length** of the request in number of bits, and
5. The **strength of the signal** received by the mobile device.

The MAC address is the primary identifier for the mobile device and has two parts. The first part is the Organisationally Unique Identifier (OUI) which provides information about the manufacturer of the device and the second part is the identifier for the device. In modern devices, to protect users' privacy, the second part of the MAC address can also be randomised and hence may not be unique to devices. When the MAC address is randomised, it is marked as such by setting a specific bit in the probe request packet as 1. Although sequence number of the packet is strictly unique to a mobile device, we hypothesize that we can use them to estimate the number of unique devices as demonstrated by [Vanhoeft et al., 2016]; where optional information present in the probe requests - Information Elements (IE) along with the sequence numbers, have been used to fingerprint the devices. This approach has become increasingly difficult as mobile phone manufacturers have severely limited the IEs present in the probe requests thus leading us to explore methods which use only the sequence numbers. This also affects the established commercial solutions using Wi-Fi probe requests such as Blix, Walkbase, Euclid Analytics, RetailNext etc. There has been another solution proposed by [Hong et al., 2018] where the authors tried to solve the similar problem using a hidden markov models based trajectory inference algorithm but the scope of this research was limited to enclosed, exit controlled public spaces such as shopping malls, railway stations, etc.

Data collection was done with the help of custom sensors built from modifying the hardware used in Smart Street Sensors [CDRC, 2016] and updating them with custom software. The sensor is essentially a Raspberry-Pi device with Wi-Fi and 3G modules. It keeps the Wi-Fi module in 'Monitor mode' and uses the open source software - Wireshark [Combs and Contributors, 2018] to passively collect all packets sent to 'broadcast', marked with type as 'management', and subtype 'probe requests'. The MAC address in these probe requests is obfuscated at the device level using a cryptographic hashing algorithm and transmitted through 3G connection to a central database via

web-sockets protocol, where it is stored in a PostgreSQL database for further analysis. The random salt used in the hashing algorithm was rotated regularly to further mitigate the risk of de-anonymisation of the hash. Though hashing cannot completely ensure anonymisation as discussed in [Demir et al., 2014], it can sufficiently obfuscate the data; which along with a secure process of data handling, gives us reasonable security. An overall schematic of the data collection and storage process is shown in Figure ???. The ground truth on the number of pedestrian footfall was recorded using a custom Android application - Clicker [Soundararaj, 2018]. This app logs accurate timestamps each time the surveyor records a pedestrian crossing the designated cordon line at the location. In addition to counting the pedestrians manually, this procedure results in the device broadcasting probe requests regularly, which in turn, gives us a ‘known device’ to calibrate our methodology against.

After collecting data, we began estimating the footfall or pedestrian activity from them by identifying the following potential uncertainties arising from our data collection method:

1. **Background noise** - since the extent to which Wi-Fi signals travel differs subject to various factors such as interference and humidity, it is close to impossible to restrict our data collection to a finite area of interest. This can lead to a significant background noise at certain locations. For example, a phone shop or a bus stop located next to the study area can artificially increase the number of probe requests received by the sensor. It is important to note that this method may not work effectively on study locations with complex configurations such as the source of noise and the area of study being located at the same distance from the sensor. This aspect is explored in detail in the broader case study in the following sections.
2. **MAC randomisation** - mobile devices in recent years have been using randomised ‘local’ MAC addresses for probe requests to protect the users from being tracked. This makes it impossible to tell if the probe requests are being sent by the same mobile device. This along with the previous problem can further increase the magnitude of error by several fold.
3. **Mobile ownership** - since the rate of mobile ownership can vary widely across geography and demography, we cannot assume that every mobile device translates to one pedestrian footfall. In addition to this, there is a long term overall increase in mobile ownership which may affect the number of probe requests collected overtime.

We propose the following internal and external validation methods to tackle each of these uncertainties.

#### *Filtering with Signal Strength*

One of the clues that we can use to estimate the distance between the mobile device and the sensor is the strength of the signal received by the sensor. The obvious approach was to first try and establish a relationship between the signal strength and distance and to use this to filter out the unwanted probe requests. However this approach was found to not be feasible, since the decay of signal strength with respect to distance is not always constant. For

instance, signal strength varies with atmospheric conditions, the presence of obstructions between the source and the target, the nature of these obstructions, and the strength (power level) of the source. This severely limits our ability to establish a simple conversion between reported signal strength and distance. As such, there was a need for a method which takes in to account all of these variables across the various locations.

We assumed that, in configurations where a specific source of background noise was at a constant distance, there must be a distinct pattern in the number of probe requests reporting signal strength corresponding to that distance. For example, if there was a phone shop next to our sensor where hundreds of phones regularly sent probe requests, there should be a sharp rise in the number of probe requests with reported signal strength corresponding to the distance between the sensor and the phone shop, irrespective of the local conditions as shown in Figure ???. We could identify these breaks in the data using traditional one dimensional clustering algorithms such as 'jenks natural breaks', 'k-means', 'quantile' and 'hierarchical clustering', etc. Since we were only looking for the break in the data and not for absolute values, the methodology should apply for all the variations due to micro site conditions reducing the overall noise in the collected data.

#### *Clustering with sequence numbers*

Since our primary unique identifier - MAC addresses are being anonymised by new devices, we needed to find other information present in the probe requests for use as a unique identifier. The obvious approach was to establish a factor of randomisation, and adjust the counts for the probe requests based on this factor. We found this approach to not be feasible since the proportion of devices which randomise the MAC addresses increased over time. There was also a wide variation in the frequency at which the devices randomised the MAC addresses and the method used for the process. This led us to look for a more generalisable approach which was independent of the device model. From our initial look at the data, we found that OUI and the sequence number of the packet was the most promising information to achieve this. First we divided our dataset into sets of probe requests with randomised and non-randomised MAC addresses by looking at the second character of the vendor part of the MAC address; if it was E, A, 2 or 6, then those addresses were identified to be randomised. We kept the MAC address as the unique identifier for the non-randomised requests and further divided the randomised ones in to sub-categories based on their OUI. We then identified unique mobile devices from within those sets, and assigned a unique identifier to each device.

The proposed algorithm created a graph where the probe requests represented the nodes; links were created between them based on the following rules:

- A link could go only forward in time.
- A link could go from low to high sequence numbers.
- A link could exist between nodes with a maximum time difference of  $\alpha$  - time threshold.

- A link could exist between nodes with a maximum sequence number difference of  $\beta$  - sequence threshold.
- A node could have only one incoming link and one outgoing link, which is the shortest of all such possible links in terms of both time and sequence number.

The nodes were then assigned a unique ID based on the unique connected component they belonged to as shown in Figure ?? . This unique identifier was used in the place of MAC addresses for aggregation of the anonymised probe requests. Although the recycling of sequence numbers after 4096 led to multiple unique IDs being reported from a single device, a sample consisting of all randomised probe requests sent by "Google" devices showed that only 0. % of the sample had their sequence number reset. This led assume this to be inconsequential.

#### *Calibrating with Ground Truth*

Since proportions of mobile device ownership was an external uncertainty to our study and could arise from variety of spatio - temporal and demographic factors, we aimed to solve this by using a manual sample count at each location. We then calculated an adjustment factor, or an 'offset' for each location by comparing the sensor-based counts and ground truth. In turn it was then used to adjust the data reliably to reflect the ground truth in absolute numbers. This calibration can be carried out periodically at these locations to improve the quality of the estimation.

#### *4.2.1 pilot study*

Table 4.7: Comparison of clustering algorithms with a sample of 40000 probe requests

Algorithm	Time (s)	MAPE (%)
Quantile	0.002	27 %
K-Means	0.007	-23 %
Hierarchical Clustering	172.520	-9 %
Bagged Clustering	0.135	-30 %
Fisher	3.034	-30 %
Jenks Natural Break	556.279	-30 %

To start, we designed a small pilot study to validate the filtering and clustering methodology against the scale and complexity of data collected in an open public area such as a retail high street. We also aimed to find the algorithm which was best suited for the classification of signal strengths as 'low' and 'high' in order to filter out the background noise. The data was collected at Oxford Street, London on 20 December 2017 from 12:30 to 13:00 hrs, Wi-Fi probe requests were collected using the sensor described in Section ?? and pedestrian footfall was manually recorded using the Android app - Clicker [Soundararaj, 2018]. Being located at one of the busiest retail locations in the United Kingdom, the Wi-Fi sensor captured approximately 60,000 probe requests during the half hour period; 3,722 people were manually recorded walking on the pavement during that time. The surveyor positioned himself at the front of a store while carrying the sensor in a backpack and counted

people walking by the store on the pavement (3m wide approximately) using a mobile phone. The sensor was kept as close to the store window as possible, and the manual count was done as a cordon count in front of the store.

As a first step we aggregated the probe requests by their MAC addresses for every minute to generate a minute by minute count of the number of people near the sensor. We assumed that each MAC address corresponded to a mobile device and hence a pedestrian. We then compared this preliminary ‘footfall’ count to the actual number of pedestrians recorded manually to check for its robustness. We used Mean Absolute Percentage Error (MAPE) as a measure of robustness of the count, since it provided a simple and quick measurement and the street conditions ensured that there are no intervals without any footfall. We found that the MAPE in the raw counts compared to the ground truth was around 425%. This suggests the presence of large amount of noise in the data which may have been generated by the sources of uncertainties discussed in Section ?? thus demonstrating the need for filtering the data.

We then classified the probe requests as ‘high signal strength’ and ‘low signal strength’ using various one dimensional clustering algorithms such as k-means, quantile, hierarchical clustering, bagged clustering, fisher and jenks natural breaks with the number of clusters set as 2. The results are shown in Table 4.7. We found that while hierarchical clustering and jenks gave us fairly low errors, they were too resource intensive for practical use with a larger dataset. We also found that k-means gave the quickest results with the lowest MAPE, closely followed by quantile algorithm. The cut-off point or threshold for the collected data with which we could classify as high and low was -71 dBm. We then removed all the probe requests which reported ‘low signal strength’ and repeated the same aggregation process as before to produce footfall count. This process resulted in a footfall count with a net MAPE of 30%. Although the results are encouraging we are still not completely confident that our filtering process is removing noise or has any correlation the configuration of sensor or position of the mobile devices. These concerns need to be addressed with a larger survey with multiple locations of varying orientations.

The next challenge was to identify the probe requests which are generated by the same device irrespective of the MAC randomisation process. We used the algorithm defined in Section ?? and assigned a unique identifier or signature to each probe request, independent of their the MAC addresses. Since we didn’t know the nature or frequency of the MAC address randomisation process, we used the surveyor’s mobile device as a reference. As the surveyor’s device was being actively used to count pedestrians and its Wi-Fi module was kept active without establishing connection to any network, the device was known to be continuously probing for new networks. We also knew that the OUI of the device was ‘Google’ and the device was regularly randomising its MAC address, thus providing us an excellent reference with which we could optimise the parameters for our clustering algorithm. We then, by increasing the thresholds in steps of 1, found the minimum possible threshold for time and sequence numbers at which the algorithm clusters the reference device properly without over clustering the other probe requests. This process is shown in Figure ???. We observed that the threshold for time

$\alpha$  and the threshold for sequence numbers,  $\beta$  are 16 seconds and 60 respectively. This was undertaken on top the filtering done based on signal strength, and only for the probe requests with randomised MAC addresses. Figure ?? shows the results of this clustering process on a small set of randomised probe requests. The probe requests with different randomised MAC address are shown by the coloured points and the lines joining them show that those probe requests were most likely be generated by the same device. We finally aggregated the probe requests as we did before but with the device signature rather than MAC addresses. This results in a footfall count with a MAPE of -18% compared to the manual count. A comparison of minute by minute counts resulting from different filtering processes along with the ground truth is shown in Figure ?? illustrating the promising effectiveness of the methods.

To conclude, from the pilot study we found that both the filtering and the clustering methods we devised worked on complex real world data and resulted in final pedestrian counts within a MAPE of 20%. We also found that ‘k-means’ and ‘quantile’ are best algorithms for clustering signal strengths. Finally, we observed that the best thresholds for time and sequence numbers in the clustering algorithm is around 16 and 60 respectively.

#### 4.2.2 main study

The methodology set out above was implemented in five different Central London locations at different times. Sensors were installed and data collected for extended periods of time. We also carried out manual counting at these locations across different times of the day. We then applied the methodologies discussed earlier to arrive at estimated pedestrian footfall and compared them with the corresponding manual counts. We finally evaluated the effectiveness of the processes with the Mean Absolute Percentage Error (MAPE) at the locations and report our findings below.

Table 4.8: Locations where sensors were installs, volume and speed of probe requests collected by the vendor and total pedestrians manually counted. The data occupies around 1.8 GB on disk when encoded in text format.

ID	Location	Type	Installation notes	Probe Requests	Footfall
1	Camden High Street	Phone Shop	Bus stop in front	9.9 (297)	3683 (33)
2	Central St.Giles	Restaurant	Seating area on both sides	3.9 (169)	0346 (05)
3	Holborn Station	Info. Kiosk	Overlooks station entrance	4.3 (303)	2956 (46)
4	Brunswick Center	Fast Food	Has seating area on one side	3.4 (210)	0960 (12)
5	The Strand	Tea Shop	Has phone shop next door	8.4 (382)	1969 (21)

The locations at which the data were collected are shown in Table 4.8. The locations were chosen for their diverse site conditions and unique sources of

noise around the potential location of the sensors. The position of the sensor at these locations with respect to the context is shown the Figure ???. We can see that Location 5 is the ‘cleanest’ with one clear stationary source of noise (phone shop) while location 2 is the most complex due to the proximity of seating areas to the sensor. The sensors were operational through out February and March, while manual counts were conducted in these locations in half hour sessions on at least two different days. For the purposes of comparing with ground truth, we considered the data from sensors which correspond to the 12 sets of available manual counts. The schedule of data collection is shown in Figure ??.

We begin by looking at the distribution of the signal strength reported by the probe requests across the locations. From the density plot shown in Figure ??, we can observe that there is clear relation between the distribution of the signal strength and the distance and complexity of the source of noise. We can see that while location 5 shows clean difference between low and high signal strengths, location 2 is almost normally distributed. Intuitively we expected that location 2 and 4 would be harder to classify than locations 1, 3 and 5. We ran the k-means clustering algorithm and filtered out the probe requests which were randomised and had signal strengths less than the second break (threshold). It is important to note that we were dealing with relative thresholds of signal strengths which can vary with location and time of the analysis. We then aggregated the probe requests by counting the number of Unique MAC addresses present in every minute. We also removed devices that dwelled around the sensor by removing the MAC addresses which reappeared within the previous hour.

The results of the first stage of the filtering process along with the thresholds are shown in Table 4.9. Confirming our intuition, we see that the location 2 has the most MAPE followed by location 4, while the rest of them have highly reduced MAPE. It is significant that this method alone reduces our margin of error by 50 - 100% from the raw counts without any cleaning. This makes the signal strength filtering a quick and ideal method for practical applications, one which doesn’t require absolute numbers such as creating large aggregated indexes to show long-term trends. We also found that the success of the signal strength filtering can be improved significantly by installing sensors so that the pedestrians and source of noise are at different distances from the sensor. This ensures that the distribution of signal strengths within the field of measurement is distinct from that of the surroundings.

We then ran the sequence numbers based clustering process on the rest of the probe requests to reduce the MAPE by almost 50 - 100% on all the sensors except for location 3. Location 3 is an outlier among all the other sensors since it is the only one with large amount of pedestrians very close to the sensor. This may be the reason behind the over filtering observed in the previous process. We finally ran the calibration process where we calculated the adjustment factors from the ratio between the manual counts to the sensor based counts for the sample period as shown in Table 4.9. We used them to adjust the counts to achieve a MAPE ranging from 10 - 50%. We observed that the sensors with people moving right next to them tend to under-count with our methodology, while sensors with seating next to them tend to over-count significantly. However, using the filtering process, we can reduce the error to

almost 10% closer to that of the ground truth.

Table 4.9: Results of footfall estimation at each location as Mean Absolute Percentage Error (MAPE) after each step of the filtering process

Sensor	Signal strength threshold	Adjustment factor	MAPE without any	MAPE after filtering signal	MAPE after filtering sequence	MAPE of final adjusted
1	-70	1.25	259	22	-13	9
2	-74	0.51	928	396	206	55
3	-72	1.60	87	-19	-31	10
4	-70	0.88	498	142	52	33
5	-72	0.80	473	84	38	11

#### 4.2.3 Conclusions

Sentient technologies make measurement of the human activities that are the life blood of the smart city possible. Yet the data that they harvest are frequently relevant only to the sub-groups within society that avail themselves of particular goods and services – such as social media applications, transport modes or retail offers. In each of these cases, it is necessary to remember that the resulting data are by-products of consumer transactions, and will as a consequence, only pertain to users of the relevant goods or services. If the smart city is to be socially inclusive, it therefore follows that sentient data must represent entire populations, whether by design or by triangulation with external, population wide, sources. This is a non-trivial task, since the ebbs and flows of smart device-enabled citizens rarely pertain to any clearly defined population in either administrative or functional terms [Massam, 1975].

Our objective here has been to collect, rather than re-use, data on smart city functioning, by recording Wi-Fi probes and ultimately reconciling them with manual counts in order to infer ambient populations. The internal validation methodology set out in the technical sections of this paper, allied to external validation from pedestrian counts, renders the method inclusive and robust when recording activity levels in retail centres in real time. We have described the collection and processing of a novel consumer Big Dataset that enables valid measures of levels of footfall activity which has been scaled across a wide network of sensors [Longley et al., 2018]. In both conceptual and technical terms, it illustrates the ways in which passively collected consumer data can be ‘hardened’ to render them robust and reliable by using related procedures of internal and external validation.

Internal validation addresses the issues of screening out device probes that do not indicate footfall, and the further screening of device probes to ‘fingerprint’ the effects of MAC randomization. It is important to note that the filtering process work based solely on the information present in the probe requests and their temporal distribution. This ensures that although the mobile devices were uniquely identified, there was no further personal data generated by linking the probe requests to the users of the mobile devices. This method essentially gave us a way to estimate the footfall in real-time without identifying or tracking the mobile devices themselves. External validation then entailed reconciling adjusted counts with the footfall observed at sample locations. This procedure makes it possible to generalise from locations at which manual footfall surveys are conducted to all others in the system, and

to develop a classification of device locations that are more or less susceptible to noise generation.

This Wi-Fi based footfall counting methodology offers a large number of applications and benefits for real time spatial analysis. Since Wi-Fi based sensors are inexpensive and the data model is scalable, it is possible to use this methodology for a large network of sensors to gather granular data on pedestrian footfall. A snapshot showing a week's worth of precise footfall in area around Charring cross, London is shown in Figure ?? in order to demonstrate the potential for such a dataset. Projects such as SmartStreetSensors [Longley et al., 2018], may utilise this methodology to overcome the challenges introduced by the implementation of MAC address randomisation.

The vicissitudes of MAC randomisation, and the provisions of privacy legislation such as EU General Data Protection Regulations mitigate against tracking individuals across the smart city using this approach. This can be modelled using agent-based methods [Heppenstall et al., 2011], however. In our own research we have also begun to link store time-lagged till receipts to footfall, and have used such data to better understand the dwell times that characterise such different retail uses as stores with window displays and fast food restaurants. Such analysis not only provides a more nuanced picture of movement through retail areas, but also enables valorisation of micro sites within retail centres. In the UK, for example, this is of immediate practical importance in evaluating business rates on properties, and has still wider implications for the setting of retail unit rental values. There are obvious extensions to understanding the ebbs and flows of activities in the 24-hour smart city such as understanding urban mobility [Gariazzo et al., 2019] and conceptualising them with a people dimension [Nam and Pardo, 2011].

More broadly still, extensions to this strand of smart city research are likely to seek to differentiate the quality of different elements within footfall according to mission e.. travel to adjacent workplace zones, leisure, etc. and personal characteristics such as spending power. In this respect, future research may not only simulate linkage of harmonised footfall counts between sensor locations, but also link these in turn to disaggregate origin-destination matrices for bikeshare and other public transport modes. Our own investigations will consider these and other challenges to understanding the functioning of the sentient city.



# 5

## *Visualisations and Applications*

5.1 *Footfall Indices*

5.2 *Events Detection*

5.3 *Pedestrian Flows*



*Discussion and Conclusions*



# 7

## Appendix

### 7.1 Benchmarking Data Tookit

This section documents all the code that has been used in the research. The programming languages used are including but not limited to R, Bash, JavaScript and SQL.

#### 7.1.1 Simple implementation in R

This R script lists all files in a given folder, parses them as JSON data serially, aggregates the records for each time interval and finally writes it to disk as a CSV file.

```
1 #! /usr/bin/Rscript
2 suppressMessages(library(tidyverse))
3 suppressMessages(library(RJSONIO))
4
5 day <- "~/unorganised-files/ff_sample/2018/01/01"
6 sensors <- paste(day, dir(day), sep = "/")[1:25]
7 probes <- NULL
8
9 for(sensor in sensors) {
10   files <- paste(sensor, dir(sensor), sep = "/");
11   for( file in files ) {
12     records <- fromJSON(file);
13     location <- vector(); signal <- vector();
14     mac <- vector(); packets <- vector();
15     vendor <- vector(); type <- vector(); time <- vector();
16     for(record in records) {
17       t <- strsplit(strsplit(file, '\\.')[[1]][1], '/')[[1]][8]
18       l <- strsplit(strsplit(file, '\\.')[[1]][1], '/')[[1]][7]
19       signal <- append(signal, record$Signal);
20       mac <- append(mac, record$MacAddress);
21       packets <- append(packets, record$PacketCount);
22       type <- append(type, record$PacketType);
23       vendor <- append(vendor, record$VendorMacPart);
24       time <- append(time, t);
25       location <- append(location, l);
```

```

26      }
27      recordsdf <- data.frame(location, time, signal,
28                                mac, packets, type, vendor);
29      if(is.null(probes)) { probes <- recordsdf; }
30      else { probes <- rbind(probes, recordsdf); }
31    }
32  }
33
34 probes %>%
35   group_by(location, time) %>%
36   summarise(count = length(unique(paste0(vendor, mac)))) %>%
37   write.csv("output-old.csv", row.names = FALSE);

```

### 7.1.2 Serial implementation in bash

This bash script lists all the files in a given folder, parses them into JSON data *serially*, aggregates the resulting records for each time interval and finally writes it to disk as a CSV file.

```

1 #! /bin/bash
2 awkc="awk -vFPAT='[^,]*|\"[^\"]*\" -v OFS=','"
3 FOLDER="/home/ucfnbso/unorganised-files/ff_sample/2018/01/01/"
4 SENSORS=`ls $FOLDER | head -n $1`
5
6 for SENSOR in $SENSORS;
7 do
8   jq_string=".[] | \
9     [\"$SENSOR\", \
10      ( (input_filename/\"/\") | .[ .length-1 ] )/\".\\" | .[0]), \
11      .VendorMacPart+.MacAddress] \
12    | @csv";
13   cmd="jq -r '$jq_string' $FOLDER$SENSOR/*.pd \
14     | sort | uniq \
15     | $awkc '{print \$1,\$2}' \
16     | sort | uniq -c";
17   echo "$(eval $cmd)" > output-new.csv;
18 done

```

### 7.1.3 Parallel implementation in bash

This bash script lists all the files in a given folder, parses them into JSON data *in parallel*, aggregates the resulting records for each time interval and finally writes it to disk as a CSV file.

```

1 #! /bin/bash
2 awkc="awk -vFPAT='[^,]*|\"[^\"]*\" -v OFS=','"
3 folder="/home/ucfnbso/unorganised-files/ff_sample/2018/01/01/"
4 sensors=`ls $folder | head -n $1`
5

```

```
6 jq_string=".[] | \  
7 [\"{}\",\  
8 ( input_filename/" / " | .[ .length-1 ] )/".\" | .[0]),\  
9 .VendorMacPart+.MacAddress] \  
10 | @csv";  
11 cmd="jq -r '$jq_string' $folder{}/*.pd \  
12 | sort | uniq \  
13 | awkc '{print \"$1,$2\"}' \  
14 | sort | uniq -c";  
15  
16 echo "$sensors" \  
17 | parallel "$cmd" \  
18 > output-new-parallel.csv
```

## 7.2 Sample Probe Request

This is a sample probe request captured using tshark and saved in the JSON format.

```

1   {
2     "_index": "packets-2017-11-15",
3     "_type": "pcap_file",
4     "_score": null,
5     "_source": {
6       "layers": {
7         "frame": {
8           "frame.interface_id": "0",
9           "frame.interface_id_tree": {
10             "frame.interface_name": "en0"
11           },
12           "frame.encap_type": "23",
13           "frame.time": "Nov 15, 2017 18:46:56.155602000 GMT",
14           "frame.offset_shift": "0.000000000",
15           "frame.time_epoch": "1510771616.155602000",
16           "frame.time_delta": "0.019159000",
17           "frame.time_delta_displayed": "0.019159000",
18           "frame.time_relative": "0.343422000",
19           "frame.number": "6",
20           "frame.len": "142",
21           "frame.cap_len": "142",
22           "frame.marked": "0",
23           "frame.ignored": "0",
24           "frame.protocols": "radiotap:wlan_radio:wlan"
25         },
26         "radiotap": {
27           "radiotap.version": "0",
28           "radiotap.pad": "0",
29           "radiotap.length": "25",
30           "radiotap.present": {
31             "radiotap.present.word": "0x00000086f",
32             "radiotap.present.word_tree": {
33               "radiotap.present.tsft": "1",
34               "radiotap.present.flags": "1",
35               "radiotap.present.rate": "1",
36               "radiotap.present.channel": "1",
37               "radiotap.present.fhss": "0",
38               "radiotap.present.dbm_antsignal": "1",
39               "radiotap.present.dbm_antnoise": "1",
40               "radiotap.present.lock_quality": "0",
41               "radiotap.present.tx_attenuation": "0",
42               "radiotap.present.db_tx_attenuation": "0",
43               "radiotap.present.dbm_tx_power": "0",
44               "radiotap.present.antenna": "1",

```

```

45     "radiotap.present.db_antsignal": "0",
46     "radiotap.present.db_antnoise": "0",
47     "radiotap.present.rxflags": "0",
48     "radiotap.present.xchannel": "0",
49     "radiotap.present.mcs": "0",
50     "radiotap.present.ampdu": "0",
51     "radiotap.present.vht": "0",
52     "radiotap.present.timestamp": "0",
53     "radiotap.present.reserved": "0x00000000",
54     "radiotap.present.rtap_ns": "0",
55     "radiotap.present.vendor_ns": "0",
56     "radiotap.present.ext": "0"
57   },
58 },
59 "radiotap.mactime": "836459236",
60 "radiotap.flags": "0x00000012",
61 "radiotap.flags_tree": {
62   "radiotap.flags.cfp": "0",
63   "radiotap.flags.preamble": "1",
64   "radiotap.flags.wep": "0",
65   "radiotap.flags.frag": "0",
66   "radiotap.flags.fcs": "1",
67   "radiotap.flags.datapad": "0",
68   "radiotap.flags.badfcs": "0",
69   "radiotap.flags.shortgi": "0"
70 },
71 "radiotap.datarate": "6",
72 "radiotap.channel.freq": "5180",
73 "radiotap.channel.flags": "0x00000140",
74 "radiotap.channel.flags_tree": {
75   "radiotap.channel.flags.turbo": "0",
76   "radiotap.channel.flags.cck": "0",
77   "radiotap.channel.flags.ofdm": "1",
78   "radiotap.channel.flags.2ghz": "0",
79   "radiotap.channel.flags.5ghz": "1",
80   "radiotap.channel.flags.passive": "0",
81   "radiotap.channel.flags.dynamic": "0",
82   "radiotap.channel.flags.gfsk": "0",
83   "radiotap.channel.flags.gsm": "0",
84   "radiotap.channel.flags.sturbo": "0",
85   "radiotap.channel.flags.half": "0",
86   "radiotap.channel.flags.quarter": "0"
87 },
88 "radiotap.dbm_antsignal": "-76",
89 "radiotap.dbm_antnoise": "-96",
90 "radiotap.antenna": "1"
91 },
92 "wlan_radio": {
93   "wlan_radio.phy": "5",

```

```

94         "wlan_radio.11a.turbo_type": "0",
95         "wlan_radio.data_rate": "6",
96         "wlan_radio.channel": "36",
97         "wlan_radio.frequency": "5180",
98         "wlan_radio.signal_dbm": "-76",
99         "wlan_radio.noise_dbm": "-96",
100        "wlan_radio.timestamp": "836459236",
101        "wlan_radio.duration": "180",
102        "wlan_radio.duration_tree": {
103            "wlan_radio.preamble": "20",
104            "wlan_radio.ifs": "19096",
105            "wlan_radio.start_tsf": "836459056",
106            "wlan_radio.end_tsf": "836459236"
107        },
108    },
109    "wlan": {
110        "wlan.fc.type_subtype": "4",
111        "wlan.fc": "0x00004000",
112        "wlan.fc_tree": {
113            "wlan.fc.version": "0",
114            "wlan.fc.type": "0",
115            "wlan.fc.subtype": "4",
116            "wlan.flags": "0x00000000",
117            "wlan.flags_tree": {
118                "wlan.fc.ds": "0x00000000",
119                "wlan.fc.tods": "0",
120                "wlan.fc.fromds": "0",
121                "wlan.fc.frag": "0",
122                "wlan.fc.retry": "0",
123                "wlan.fc.pwrctrl": "0",
124                "wlan.fc.moredata": "0",
125                "wlan.fc.protected": "0",
126                "wlan.fc.order": "0"
127            },
128        },
129        "wlan.duration": "0",
130        "wlan.ra": "ff:ff:ff:ff:ff:ff",
131        "wlan.ra_resolved": "Broadcast",
132        "wlan.da": "ff:ff:ff:ff:ff:ff",
133        "wlan.da_resolved": "Broadcast",
134        "wlan.ta": "94:b1:0a:79:15:9b",
135        "wlan.ta_resolved": "SamsungE_79:15:9b",
136        "wlan.sa": "94:b1:0a:79:15:9b",
137        "wlan.sa_resolved": "SamsungE_79:15:9b",
138        "wlan.bssid": "ff:ff:ff:ff:ff:ff",
139        "wlan.bssid_resolved": "Broadcast",
140        "wlan.addr": "ff:ff:ff:ff:ff:ff",
141        "wlan.addr_resolved": "Broadcast",
142        "wlan.addr": "94:b1:0a:79:15:9b",

```

```

143     "wlan.addr_resolved": "SamsungE_79:15:9b",
144     "wlan.addr": "ff:ff:ff:ff:ff:ff",
145     "wlan.addr_resolved": "Broadcast",
146     "wlan.frag": "0",
147     "wlan.seq": "2533",
148     "wlan.fcs": "0x5a6629b7",
149     "wlan.fcs.status": "1"
150   },
151   "wlan": {
152     "wlan.tagged.all": {
153       "wlan.tag": {
154         "wlan.tag.number": "0",
155         "wlan.tag.length": "9",
156         "wlan.ssid": "VM3280449"
157       },
158       "wlan.tag": {
159         "wlan.tag.number": "1",
160         "wlan.tag.length": "8",
161         "wlan.supported_rates": "12",
162         "wlan.supported_rates": "18",
163         "wlan.supported_rates": "24",
164         "wlan.supported_rates": "36",
165         "wlan.supported_rates": "48",
166         "wlan.supported_rates": "72",
167         "wlan.supported_rates": "96",
168         "wlan.supported_rates": "108"
169       },
170       "wlan.tag": {
171         "wlan.tag.number": "45",
172         "wlan.tag.length": "26",
173         "wlan.ht.capabilities": "0x00000062",
174         "wlan.ht.capabilities_tree": {
175           "wlan.ht.capabilities.ldpcoding": "0",
176           "wlan.ht.capabilities.width": "1",
177           "wlan.ht.capabilities.sm": "0x00000000",
178           "wlan.ht.capabilities.green": "0",
179           "wlan.ht.capabilities.short20": "1",
180           "wlan.ht.capabilities.short40": "1",
181           "wlan.ht.capabilities.txstbc": "0",
182           "wlan.ht.capabilities.rxstbc": "0x00000000",
183           "wlan.ht.capabilities.delayedblockack": "0",
184           "wlan.ht.capabilities.amsdu": "0",
185           "wlan.ht.capabilities.dsscck": "0",
186           "wlan.ht.capabilities.psmp": "0",
187           "wlan.ht.capabilities.40mhzintolerant": "0",
188           "wlan.ht.capabilities.lsig": "0"
189         },
190         "wlan.ht.ampduparam": "0x00000017",
191         "wlan.ht.ampduparam_tree": {

```

```

192         "wlan.ht.ampdparam.maxlength": "0x00000003",
193         "wlan.ht.ampdparam.mpudensity": "0x00000005",
194         "wlan.ht.ampdparam.reserved": "0x00000000"
195     },
196     "wlan.ht.mcsset": "MCS Set",
197     "wlan.ht.mcsset_tree": {
198         "wlan.ht.mcsset.rxbitmask": {
199             "wlan.ht.mcsset.rxbitmask.0to7": "0x000000ff",
200             "wlan.ht.mcsset.rxbitmask.8to15": "0x00000000",
201             "wlan.ht.mcsset.rxbitmask.16to23": "0x00000000",
202             "wlan.ht.mcsset.rxbitmask.24to31": "0x00000000",
203             "wlan.ht.mcsset.rxbitmask.32": "0x00000001",
204             "wlan.ht.mcsset.rxbitmask.33to38": "0x00000000",
205             "wlan.ht.mcsset.rxbitmask.39to52": "0x00000000",
206             "wlan.ht.mcsset.rxbitmask.53to76": "0x00000000"
207         },
208         "wlan.ht.mcsset.highestdatarate": "0x00000000",
209         "wlan.ht.mcsset.txsetdefined": "0",
210         "wlan.ht.mcsset.txrxmcsnotequal": "0",
211         "wlan.ht.mcsset.txmaxss": "0x00000000",
212         "wlan.ht.mcsset.txunequalmod": "0"
213     },
214     "wlan.htex.capabilities": "0x00000000",
215     "wlan.htex.capabilities_tree": {
216         "wlan.htex.capabilities.pco": "0",
217         "wlan.htex.capabilities.transtime": "0x00000000",
218         "wlan.htex.capabilities.mcs": "0x00000000",
219         "wlan.htex.capabilities.htc": "0",
220         "wlan.htex.capabilities.rdresponder": "0"
221     },
222     "wlan.txbf": "0x00000000",
223     "wlan.txbf_tree": {
224         "wlan.txbf.txbf": "0",
225         "wlan.txbf.rxss": "0",
226         "wlan.txbf.txss": "0",
227         "wlan.txbf.rxndp": "0",
228         "wlan.txbf.txndp": "0",
229         "wlan.txbf.impltxbf": "0",
230         "wlan.txbf.calibration": "0x00000000",
231         "wlan.txbf.csi": "0",
232         "wlan.txbf.fm.uncompressed.tbf": "0",
233         "wlan.txbf.fm.compressed.tbf": "0",
234         "wlan.txbf.rcsi": "0x00000000",
235         "wlan.txbf.fm.uncompressed.rbf": "0x00000000",
236         "wlan.txbf.fm.compressed.bf": "0x00000000",
237         "wlan.txbf.mingroup": "0x00000000",
238         "wlan.txbf.csinumant": "0x00000000",
239         "wlan.txbf.fm.uncompressed.maxant": "0x00000000",
240         "wlan.txbf.fm.compressed.maxant": "0x00000000",

```

```

241     "wlan.txbf.csi.maxrows": "0x00000000",
242     "wlan.txbf.channelest": "0x00000000",
243     "wlan.txbf.reserved": "0x00000000"
244   },
245   "wlan.asel": "0x00000000",
246   "wlan.asel_tree": {
247     "wlan.asel.capable": "0",
248     "wlan.asel.txcsi": "0",
249     "wlan.asel.txif": "0",
250     "wlan.asel.csi": "0",
251     "wlan.asel.if": "0",
252     "wlan.asel.rx": "0",
253     "wlan.asel.sppdu": "0",
254     "wlan.asel.reserved": "0x00000000"
255   }
256 },
257 "wlan.tag": {
258   "wlan.tag.number": "127",
259   "wlan.tag.length": "8",
260   "wlan.extcap": "0x00000000",
261   "wlan.extcap_tree": {
262     "wlan.extcap.b0": "0",
263     "wlan.extcap.b1": "0",
264     "wlan.extcap.b2": "0",
265     "wlan.extcap.b3": "0",
266     "wlan.extcap.b4": "0",
267     "wlan.extcap.b5": "0x00000000",
268     "wlan.extcap.b6": "0",
269     "wlan.extcap.b7": "0"
270   },
271   "wlan.extcap": "0x00000000",
272   "wlan.extcap_tree": {
273     "wlan.extcap.b8": "0",
274     "wlan.extcap.b9": "0",
275     "wlan.extcap.b10": "0",
276     "wlan.extcap.b11": "0",
277     "wlan.extcap.b12": "0",
278     "wlan.extcap.b13": "0",
279     "wlan.extcap.b14": "0",
280     "wlan.extcap.b15": "0"
281   },
282   "wlan.extcap": "0x00000008",
283   "wlan.extcap_tree": {
284     "wlan.extcap.b16": "0",
285     "wlan.extcap.b17": "0",
286     "wlan.extcap.b18": "0",
287     "wlan.extcap.b19": "1",
288     "wlan.extcap.b20": "0",
289     "wlan.extcap.b21": "0",

```

```

290         "wlan.extcap.b22": "0",
291         "wlan.extcap.b23": "0"
292     },
293     "wlan.extcap": "0x00000080",
294     "wlan.extcap_tree": {
295         "wlan.extcap.b24": "0",
296         "wlan.extcap.b25": "0",
297         "wlan.extcap.b26": "0",
298         "wlan.extcap.b27": "0",
299         "wlan.extcap.b28": "0",
300         "wlan.extcap.b29": "0",
301         "wlan.extcap.b30": "0",
302         "wlan.extcap.b31": "1"
303     },
304     "wlan.extcap": "0x00000001",
305     "wlan.extcap_tree": {
306         "wlan.extcap.b32": "1",
307         "wlan.extcap.b33": "0",
308         "wlan.extcap.b34": "0",
309         "wlan.extcap.b35": "0x00000000",
310         "wlan.extcap.b36": "0",
311         "wlan.extcap.b37": "0",
312         "wlan.extcap.b38": "0",
313         "wlan.extcap.b39": "0"
314     },
315     "wlan.extcap": "0x00000040",
316     "wlan.extcap_tree": {
317         "wlan.extcap.b40": "0",
318         "wlan.extcap.serv_int_granularity": "0",
319         "wlan.extcap.b44": "0",
320         "wlan.extcap.b45": "0",
321         "wlan.extcap.b46": "1",
322         "wlan.extcap.b47": "0x00000000"
323     },
324     "wlan.extcap": "0x00000000",
325     "wlan.extcap_tree": {
326         "wlan.extcap.b48": "0",
327         "wlan.extcap.o7": "0x00000000"
328     },
329     "wlan.extcap": "0x00000040",
330     "wlan.extcap_tree": {
331         "wlan.extcap.b61": "0",
332         "wlan.extcap.b62": "1",
333         "wlan.extcap.b63": "0",
334         "wlan.extcap.o8": "0x00000000"
335     }
336 },
337 "wlan.tag": {
338     "wlan.tag.number": "107",

```

```
339     "wlan.tag.length": "1",
340     "wlan.interworking.access_network_type": "15",
341     "wlan.interworking.internet": "0",
342     "wlan.interworking.asra": "0",
343     "wlan.interworking.esr": "0",
344     "wlan.interworking.uesa": "0"
345   },
346   "wlan.tag": {
347     "wlan.tag.number": "221",
348     "wlan.tag.length": "5",
349     "wlan.tag.oui": "5271450",
350     "wlan.tag.vendor.oui.type": "16",
351     "wlan.hs20.indication.dgaf_disabled": "0",
352     "wlan.hs20.indication.pps_mo_id_present": "0",
353     "wlan.hs20.indication.anqp_domain_id_present": "0",
354     "wlan.hs20.indication.release_number": "1"
355   },
356   "wlan.tag": {
357     "wlan.tag.number": "221",
358     "wlan.tag.length": "7",
359     "wlan.tag.oui": "20722",
360     "wlan.tag.vendor.oui.type": "8",
361     "wlan.wfa.ie.type": "0x00000008"
362   },
363   "wlan.tag": {
364     "wlan.tag.number": "221",
365     "wlan.tag.length": "9",
366     "wlan.tag.oui": "4120",
367     "wlan.tag.vendor.oui.type": "2",
368     "wlan.tag.vendor.data": "02:00:00:10:00:00"
369   }
370 }
```







## Estimating real-time high-street footfall from Wi-Fi probe requests

Balamurugan Soundararaj , James Cheshire and Paul Longley

Department of Geography, University College London, London, UK

### ABSTRACT

The accurate measurement of human activity with high spatial and temporal granularity is crucial for understanding the structure and function of the built environment. With increasing mobile ownership, the Wi-Fi ‘probe requests’ generated by mobile devices can act as a cheap, scalable and real-time source of data for establishing such measures. The two major challenges we face in using these probe requests for estimating human activity are: filtering the noise generated by the uncertain field of measurement and clustering anonymised probe requests generated by the same devices together without compromising the privacy of the users. In this paper, we demonstrate that we can overcome these challenges by using class intervals and a novel graph-based technique for filtering and clustering the probe requests which in turn, enables us to reliably measure real-time pedestrian footfall at retail high streets.

### ARTICLE HISTORY

Received 1 July 2018

Accepted 22 February 2019

### KEYWORDS

Pedestrian footfall; urban sensing; Wi-Fi probe requests; MAC randomisation

## 1. Introduction

New and developing ‘smart’ technologies today provide the infrastructure over which movements and interactions of people can be measured and monitored in the ‘sentient city’ (Amin and Thrift 2017). This is making it possible to reinvigorate conceptualisations of city as the locus of human activities supplementing night-time geographies of residence (Martin *et al.* 2015) with geographies of shopping behaviour (Lloyd and Cheshire 2018), workzone geographies (Singleton and Longley 2015) and studies of movement trajectories (Campbell *et al.* 2008). This is rendering activity-based conceptions of human behaviour central to the analysis of hardship and opportunity in, and around, the smart city (Venerandi *et al.* 2015).

Sentient technologies include mobile phone networks, which can triangulate user locations relative to networks of masts, the use of GPS to locate users of social media services, and Wi-Fi access points providing internet access. These technologies offer differing levels of spatial precision, where mobile telephony and Wi-Fi are generally less reliable and offer lower precision than GPS to the end users, whilst simultaneously being more advantageous for broader mobility studies (Pinelli *et al.* 2015). There has been considerable research into the utility of these technologies to understand patterns of movement in cities in near real-time (Candia *et al.* 2008, Gonzalez *et al.* 2008, Calabrese *et al.* 2013). Most of this research has focused upon technical specification of accuracy or

precision (Lane *et al.* 2010, Song *et al.* 2010), with somewhat less attention devoted to the ways that the characteristics of the technologies and of their human users, conspire to create possible bias in representing usage patterns across the entire smart city.

Analysis of mobile phone data, usually derived from industry players that have significant market share and user bases representative of local populations, may also exclude groups such as tourists from distant origins or subscribers to third-party services that share distinctive characteristics (Di Luzio *et al.* 2016).

These examples illustrate the issues that underpin the assembly and analysis of consumer data. Consumer data can be considered as a distinctive class of Big Data which arise from the interactions between humans and customer-facing organisations such as retailers, domestic energy suppliers, transport providers and suppliers of social media and communications (Longley *et al.* 2018). Consumer data account for an ever-increasing real share of all of the data that are collected about citizens, but a fundamental characteristic of consumer-led markets is that no single provider has a monopoly in market provision; therefore issues of market share, and segment generate bias in the analysis. The source and operation of this bias are unknown in the absence of extensive and context-sensitive attempts to triangulate consumer data with data of known provenance relating to clearly defined populations (Lansley and Longley 2016). In similar ways to other classes of Big Data, consumer data are best thought of as digital ‘exhaust’, or a by-product created by, or harvested from, consumer transactions.

In this paper, through a set of experiments, we evaluate the value of data collected a network of 800 devices (CDRC 2016) installed across Great Britain in order to characterize the footfall patterns of a scientifically balanced sample of retail centres. These devices are located in shop windows, and record the probes emitted by mobile phones and other Wi-Fi enabled devices. The data collected from these devices are deemed to be consumer data because devices carried by consumers routinely probe for Wi-Fi connection which is a consumer service. Monitoring the probes from such devices provides an indication of the presence of their users, regardless of whether or not internet connectivity is established. Our core motivation is to appraise the usefulness of Wi-Fi probe requests harvested from our network of sensors as a method of indicating levels of pedestrian activity. More broadly still, in our future research, we intend to classify the nationwide network of footfall profiles as part of a programme of research to understand the form and function of retail areas at a time of far-reaching structural change for the retail industry.

To this end, it was important to first undertake a thorough conceptual and technical appraisal of our consumer data source. In technical terms, screening the information present in the ‘probe requests’ and clustering them based on their characteristics was essential in order to remove those emitted by devices which do not indicate pedestrian activity, such as network-enabled printers and other fixed devices. Related to this, a method to fingerprint Wi-Fi probes was necessary in order to remove probes from individuals’ devices that in conceptual terms should not be considered part of footfall; for instance when, an employee is seated in an office within range of the sensor device. A calibration of sensor measurement was also essential on two grounds: first individuals may carry multiple devices, or no device at all; and second, the positioning and orientation of the sensor in the retail unit may lead to systematic over- or under-enumeration. These sources of bias in measurement must be accommodated by

a manual recording of footfall at each location, and the generalization of these sample survey results to all locations and time periods. As we describe in detail below, manual validation of the data needed to be undertaken in parallel with the technical profiling of the mix of consumer mobile devices that probed our sensors, since the effectiveness of data cleaning procedures discussed in this paper differ between individual locations and configurations.

## 2. Background

In the past decade, Wi-Fi has emerged as one of the most commonly used technologies in providing high-speed internet access to mobile devices such as smartphones, tablets and laptops in public and private spaces (Torrens 2008). This has resulted in multiple Wi-Fi networks being available at almost every location in dense urban environments. Traversing through this overlapping mesh of Wi-Fi networks, modern mobile devices with Wi-Fi network interfaces regularly broadcast a special type of signal known as 'Probe Requests' in order to discover the Wi-Fi networks available to them. This helps these devices to connect and switch between the Wi-Fi networks seamlessly.

Probe requests are low-level signals standardised by IEEE 802.11 specification (IEEE 2016) for service discovery, and are implemented in any Wi-Fi capable device irrespective of the manufacturer or the model. This ubiquity and standardisation make them an excellent source of open, passive, continuous, and wireless data generated by Wi-Fi capable devices present at any given time and location. Considering the unprecedented levels of mobile device ownership in recent years, we can, in turn, use this data to understand the population distribution in highly dynamic urban environments with high spatial and temporal granularity (Freudiger 2015, Kontokosta and Johnson 2017). While a Wi-Fi-based method to collect data offers us various advantages such as easy scalability and efficiency in terms of cost and time, it also introduces few systematic biases and uncertainties in the collected data along with the serious risk of infringing on the privacy of the mobile users. In this paper, using a set of probe requests and manual counts collected at various high street locations across London, we demonstrate that pedestrian footfall at these locations can be estimated with considerable precision and accuracy while protecting the privacy of the pedestrians.

Unlike GPS, the location of the Wi-Fi enabled mobile device cannot be directly inferred from Wi-Fi, however there are reliable methods to triangulate the location of mobile devices from the locations of known access points (AP) and the signal strength reported by them (He *et al.* 2003, Moore *et al.* 2004, LaMarca *et al.* 2005). This can overcome the usual shortcoming of GPS, which struggles for precision and accuracy in indoor and densely built environments (Zarimpas *et al.* 2006, Kawaguchi 2009, Xi *et al.* 2010). Utilising this, we can easily and quickly estimate trajectories of the mobile devices (Musa and Eriksson 2012) which can be used similarly to the GPS trajectories to understand individual travel patterns (Rekimoto *et al.* 2007, Sapiezynski *et al.* 2015), crowd behaviour (Abedi *et al.* 2013, Mowafi *et al.* 2013), vehicular (Lu *et al.* 2010) and pedestrian movement (Xu *et al.* 2013, Fukuzaki *et al.* 2014, Wang *et al.* 2016). Such data can also be used in transportation planning and management to estimate travel time (Musa and Eriksson 2011) and real-time traffic monitoring (Abbott-Jard *et al.* 2013). Using techniques demonstrated by Franklin *et al.* (2006) and Pang *et al.* (2007), along with

information present in the probe requests, one can even model interactions between the users (Cheng *et al.* 2012, Barbera *et al.* 2013, Cunche 2014, Cunche *et al.* 2014) such as predicting which of them are most likely to meet again (Cunche *et al.* 2012). Using the semantic information present in these probe requests it even is possible to understand the nature of the population at a large scale (Di Luzio *et al.* 2016).

Although extensive research has been carried out on this subject with feasible and favourable results, in recent years, one of the major challenges faced in such attempts has been the increasing attempt by mobile phone manufacturers to protect their users' privacy by anonymising the globally identifiable portion of the probe requests (Greenstein *et al.* 2008). Various methods have been devised to overcome this anonymisation process such as estimating the device model information from a known dataset of manufacturers and device behaviours (Martin *et al.* 2016); Scrambler attack using a small part of the physical layer specification for Wi-Fi (Bloessl *et al.* 2015, Vo-Huu *et al.* 2016); and timing attack where the packet sequence information along with information elements present in the probe request frame is used (Cheng and Wang 2016, Matte *et al.* 2016). A combination of these methodologies has been proven to produce de-anonymised globally unique device information (Vanhoeef *et al.* 2016, Martin *et al.* 2017). These approaches usually result in serious risk of breach of privacy of the users of the mobile devices by revealing their identifiable personal information.

There is a clear gap in the research for exploring methodologies for estimating the number of unique mobile devices from a set of anonymised probe requests, without the need to reveal their original device information. Such a technique has various applications such as uncovering the urban wireless landscape (Rose and Welsh 2010), revealing human activity at large scales (Qin *et al.* 2013), estimating pedestrian numbers in crowds (Schauer *et al.* 2014, Fukuzaki *et al.* 2015), and even counting people in hyperlocal scales such as queues (Wang *et al.* 2013). With enough infrastructure to collect such information, we can even aim to generate a real-time census of the city (Kontokosta and Johnson 2017). With this background, we set out to devise and implement a methodology to reliably estimate human activity such as pedestrian footfall from Wi-Fi probe requests without risking a breach of privacy of the users involved.

### 3. Methodology

The primary aim of this research was to enable us to collect a series of probe requests and process them into a usable pedestrian footfall count. We did this by using a Wi-Fi receiver to collect probe requests broadcast by mobile devices, filtering out the background noise, and aggregating them based on the device that generated them. In this section, we examine the characteristics of probe requests in detail, devise a methodology to collect these probe requests in public areas, examine the systemic biases and uncertainties in the data collection method, and devise data processing methods to overcome these challenges. Finally, we compare the processed footfall counts to the ground truth recorded by primary surveys.

Probe requests are a special type of management packet broadcast by Wi-Fi enabled devices as part of their various functions such as scanning for available APs and quick geolocation by triangulation based known APs, etc. These are broadcast by all Wi-Fi enabled devices regardless of the manufacturer, type or model of the devices, although

there is some variation in the frequency and the content of the information transmitted through them. In some cases, such as Android devices, these are broadcast even when the Wi-Fi functionality has been turned off by the user so that the device can immediately connect to networks when the functionality is switched back on. Since some devices even use the probe requests as a less accurate form of localisation, they continuously send probe requests when Wi-Fi has been switched off. Thus, these signals can be used to reliably identify the presence of Wi-Fi enabled mobile devices. Being the first step of the connection initiated by the mobile device, these packets have information regarding the characteristics of the mobile device itself. Some of the key information we can infer from these requests are,

- (1) **Media Access Control (MAC) address** which is a name identifying the wireless hardware of the mobile device,
- (2) **Sequence number** of the request for the mobile device to keep track of the responses,
- (3) **Time stamp** at which the request was received by the AP,
- (4) **Total length** of the request in number of bits, and
- (5) **The strength of the signal** received by the mobile device.

The MAC address is the primary identifier for the mobile device and has two parts. The first part is the Organisational Unique Identifier (OUI) which provides information about the manufacturer of the device and the second part is the identifier for the device. In modern devices, to protect users' privacy, the second part of the MAC address can also be randomised and hence may not be unique to devices. When the MAC address is randomised, it is marked as such by setting a specific bit in the probe request packet as 1. Although sequence number of the packet is strictly unique to a mobile device, we hypothesize that we can use them to estimate the number of unique devices as demonstrated by (Vanhoef *et al.* 2016); where optional information present in the probe requests – Information Elements (IE) along with the sequence numbers, have been used to fingerprint the devices. This approach has become increasingly difficult as mobile phone manufacturers have severely limited the IEs present in the probe requests thus leading us to explore methods which use only the sequence numbers. This also affects the established commercial solutions using Wi-Fi probe requests such as Blix, Walkbase, Euclid Analytics, RetailNext, etc.. There has been another solution proposed by (Hong *et al.* 2018) where the authors tried to solve the similar problem using a hidden Markov models based trajectory inference algorithm but the scope of this research was limited to enclose, exit controlled public spaces such as shopping malls, and railway stations.

Data collection was done with the help of custom sensors built from modifying the hardware used in Smart Street Sensors (CDRC 2016) and updating them with custom software. The sensor is essentially a Raspberry-Pi device with Wi-Fi and 3G modules. It keeps the Wi-Fi module in 'Monitor mode' and uses the open source software – Wireshark (Combs 2018) to passively collect all packets sent to 'broadcast', marked with type as 'management', and subtype 'probe requests'. The MAC address in these probe requests is obfuscated at the device level using a cryptographic hashing algorithm and transmitted through 3G connection to a central database via web-sockets protocol, where it is stored in a PostgreSQL database for further analysis. The random salt used in the hashing algorithm was rotated regularly to further mitigate the risk of de-anonymization of the hash. Though hashing cannot completely

ensure anonymisation as discussed in (Demir *et al.* 2014), it can sufficiently obfuscate the data; which along with a secure process of data handling, gives us reasonable security. An overall schematic of the data collection and storage process is shown in Figure 1. The ground truth on the number of pedestrian footfall was recorded using a custom Android application – Clicker (Soundararaj 2018). This app logs accurate timestamps each time the surveyor records a pedestrian crossing the designated cordon line at the location. In addition to counting the pedestrians manually, this procedure results in the device broadcasting probe requests regularly, which in turn, gives us a ‘known device’ to calibrate our methodology against.

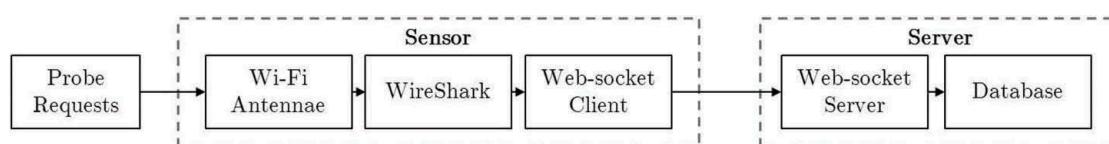
After collecting data, we began estimating the footfall or pedestrian activity from them by identifying the following potential uncertainties arising from our data collection method:

- (1) **Background noise** - since the extent to which Wi-Fi signals travel differs subject to various factors such as interference and humidity, it is close to impossible to restrict our data collection to a finite area of interest. This can lead to a significant background noise at certain locations. For example, a phone shop or a bus stop located next to the study area can artificially increase the number of probe requests received by the sensor. It is important to note that this method may not work effectively on study locations with complex configurations such as the source of noise and the area of study being located at the same distance from the sensor. This aspect is explored in detail in the broader case study in the following sections.
- (2) **MAC randomisation** - mobile devices in recent years have been using randomised ‘local’ MAC addresses for probe requests to protect the users from being tracked. This makes it impossible to tell if the probe requests are being sent by the same mobile device. This along with the previous problem can further increase the magnitude of error by several fold.
- (3) **Mobile ownership** - since the rate of mobile ownership can vary widely across geography and demography, we cannot assume that every mobile device translates to one pedestrian footfall. In addition to this, there is a long-term overall increase in mobile ownership which may affect the number of probe requests collected over time.

We propose the following internal and external validation methods to tackle each of these uncertainties.

### **3.1. Filtering with signal strength**

One of the clues that we can use to estimate the distance between the mobile device and the sensor is the strength of the signal received by the sensor. The obvious approach was to first try and establish a relationship between the signal strength and distance and to use this to filter out the unwanted probe requests. However, this



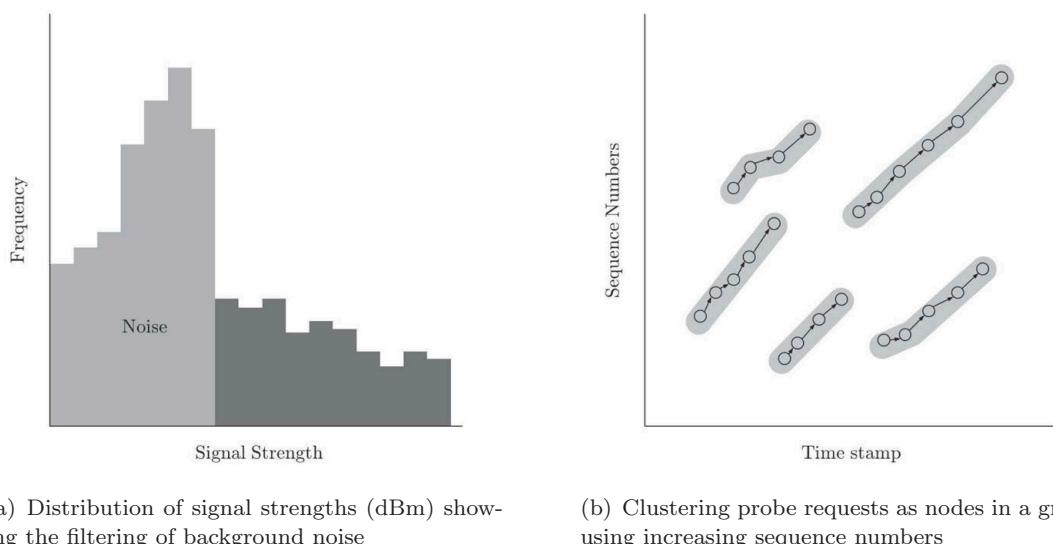
**Figure 1.** Schematic diagram showing the process of collecting and storing probe requests using the sensor.

approach was found to not be feasible since the decay of signal strength with respect to distance is not always constant. For instance, signal strength varies with atmospheric conditions, the presence of obstructions between the source and the target, the nature of these obstructions, and the strength (power level) of the source. This severely limits our ability to establish a simple conversion between reported signal strength and distance. As such, there was a need for a method which takes in to account all of these variables across the various locations.

We assumed that, in configurations where a specific source of background noise was at a constant distance, there must be a distinct pattern in the number of probe requests reporting signal strength corresponding to that distance. For example, if there was a phone shop next to our sensor where hundreds of phones regularly sent probe requests, there should be a sharp rise in the number of probe requests with reported signal strength corresponding to the distance between the sensor and the phone shop, irrespective of the local conditions as shown in Figure 2. We could identify these breaks in the data using traditional one-dimensional clustering algorithms such as 'jenks natural breaks', 'k-means', 'quantile' and 'hierarchical clustering', etc. Since we were only looking for the break in the data and not for absolute values, the methodology should apply for all the variations due to micro site conditions reducing the overall noise in the collected data.

### **3.2. Clustering with sequence numbers**

Since our primary unique identifier – MAC addresses are being anonymised by new devices, we needed to find other information present in the probe requests for use as a unique identifier. The obvious approach was to establish a factor of randomisation and adjust the counts for the probe requests based on this factor. We found this approach to not be feasible since the proportion of devices which randomise the MAC addresses increased over time. There was also a wide variation in the frequency at which the devices randomised the



**Figure 2.** Schematic diagrams explaining the methods for filtering by signal strength and clustering using sequence numbers.

MAC addresses and the method used for the process. This led us to look for a more generalisable approach which was independent of the device model.

From our initial look at the data, we found that OUI and the sequence number of the packet was the most promising information to achieve this. First, we divided our dataset into sets of probe requests with randomised and non-randomised MAC addresses by looking at the second character of the vendor part of the MAC address, if it was E, A, 2 or 6, then those addresses were identified to be randomised. We kept the MAC address as the unique identifier for the non-randomised requests and further divided the randomised ones into sub-categories based on their OUI. We then identified unique mobile devices from within those sets and assigned a unique identifier to each device.

The proposed algorithm created a graph where the probe requests represented the nodes; links were created between them based on the following rules:

- A link could go only forward in time.
- A link could go from low to high sequence numbers.
- A link could exist between nodes with a maximum time difference of  $\alpha$  – time threshold.
- A link could exist between nodes with a maximum sequence number difference of  $\beta$  – sequence threshold.
- A node could have only one incoming link and one outgoing link, which is the shortest of all such possible links in terms of both time and sequence number.

The nodes were then assigned a unique ID based on the unique connected component they belonged to as shown in [Figure 2](#). This unique identifier was used in the place of MAC addresses for aggregation of the anonymised probe requests. Although the recycling of sequence numbers after 4096 led to multiple unique IDs being reported from a single device, a sample consisting of all randomised probe requests sent by ‘Google’ devices showed that only 0.5% of the sample had their sequence number reset. This led to assume this to be inconsequential.

### **3.3. Calibrating with ground truth**

Since proportions of mobile device ownership was an external uncertainty to our study and could arise from a variety of spatio-temporal and demographic factors, we aimed to solve this by using a manual sample count at each location. We then calculated an adjustment factor or an ‘offset’ for each location by comparing the sensor-based counts and ground truth. In turn, it was then used to adjust the data reliably to reflect the ground truth in absolute numbers. This calibration can be carried out periodically at these locations to improve the quality of the estimation.

## **4. Pilot study**

To start, we designed a small pilot study to validate the filtering and clustering methodology against the scale and complexity of data collected in an open public area such as a retail high street. We also aimed to find the algorithm which was best suited for the classification of signal strengths as ‘low’ and ‘high’ in order to filter out the background noise. The data was collected

at Oxford Street, London on 20 December 2017 from 12:30 to 13:00 h, Wi-Fi probe requests were collected using the sensor described in [Section 3](#), and pedestrian footfall was manually recorded using the Android app – Clicker ([Soundararaj 2018](#)). Being located at one of the busiest retail locations in the United Kingdom, the Wi-Fi sensor captured approximately 60,000 probe requests during the half-hour period; 3,722 people were manually recorded walking on the pavement during that time. The surveyor positioned himself at the front of a store while carrying the sensor in a backpack and counted people walking by the store on the pavement (3 m wide approximately) using a mobile phone. The sensor was kept as close to the store window as possible, and the manual count was done as a cordon count in front of the store.

As a first step, we aggregated the probe requests by their MAC addresses for every minute to generate a minute by minute count of the number of people near the sensor. We assumed that each MAC address corresponded to a mobile device and hence a pedestrian. We then compared this preliminary ‘footfall’ count to the actual number of pedestrians recorded manually to check for its robustness. We used Mean Absolute Percentage Error (MAPE) as a measure of robustness of the count since it provided a simple and quick measurement and the street conditions ensured that there are no intervals without any footfall. We found that the MAPE in the raw counts compared to the ground truth was around 425%. This suggests the presence of large amount of noise in the data which may have been generated by the sources of uncertainties discussed in [Section 3](#) thus demonstrating the need for filtering the data.

We then classified the probe requests as ‘high signal strength’ and ‘low signal strength’ using various one-dimensional clustering algorithms such as k-means, quantile, hierarchical clustering, bagged clustering, fisher and jenks natural breaks with the number of clusters set as 2. The results are shown in [Table 1](#). We found that while hierarchical clustering and jenks gave us fairly low errors, they were too resource intensive for practical use with a larger dataset. We also found that k-means gave the quickest results with the lowest MAPE, closely followed by quantile algorithm. The cut-off point or threshold for the collected data with which we could classify as high and low was  $-71 \text{ dBm}$ . We then removed all the probe requests which reported ‘low signal strength’ and repeated the same aggregation process as before to produce footfall count. This process resulted in a footfall count with a net MAPE of 30%. Although the results are encouraging we are still not completely confident that our filtering process is removing noise or has any correlation the configuration of sensor or position of the mobile devices. These concerns need to be addressed with a larger survey with multiple locations of varying orientations.

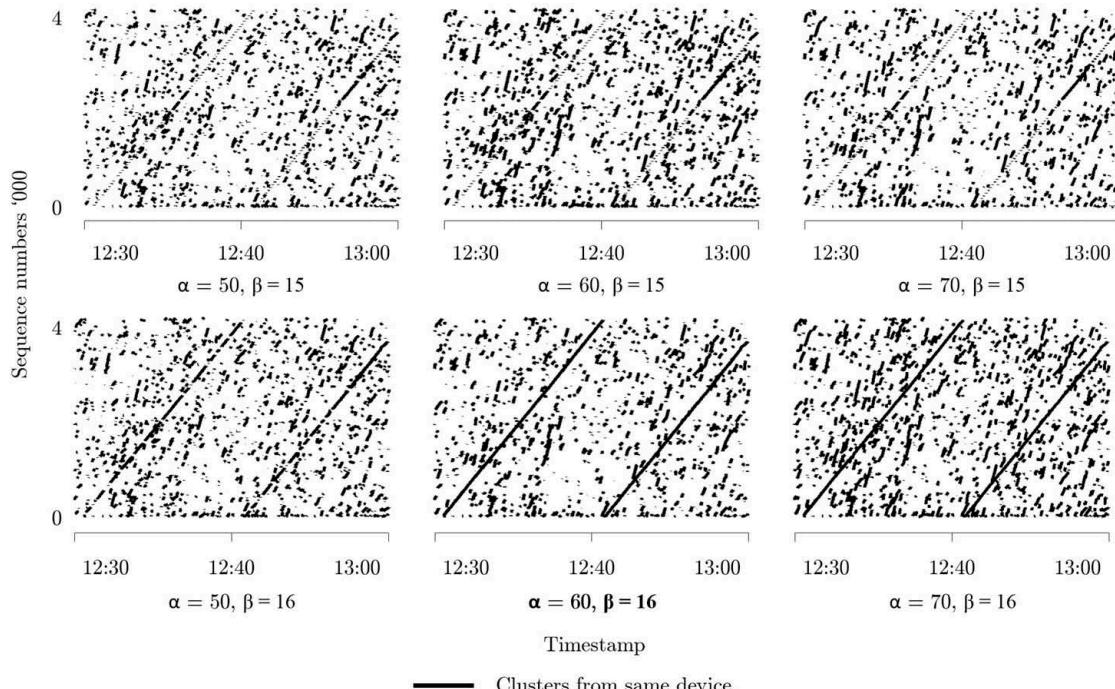
The next challenge was to identify the probe requests which are generated by the same device irrespective of the MAC randomisation process. We used the algorithm defined in [Section 3](#) and assigned a unique identifier or signature to each probe request,

**Table 1.** Comparison of clustering algorithms with a sample of 40,000 probe requests.

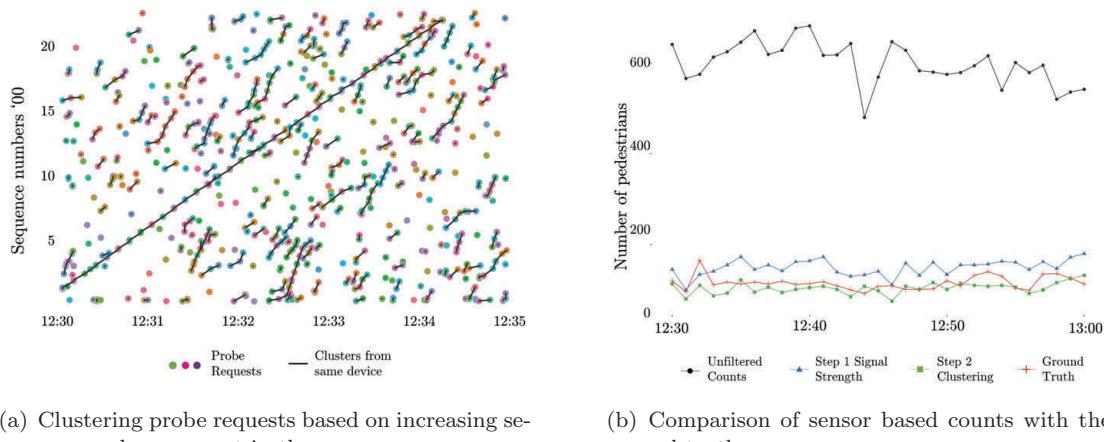
Algorithm	Time (s)	MAPE (%)
Quantile	0.002	27%
K-Means	0.007	-23%
Hierarchical Clustering	172.520	-9%
Bagged Clustering	0.135	-30%
Fisher	3.034	-30%
Jenks Natural Break	556.279	-30%

independent of their the MAC addresses. Since we did not know the nature or frequency of the MAC address randomisation process, we used the surveyor's mobile device as a reference. As the surveyor's device was being actively used to count pedestrians and its Wi-Fi module was kept active without establishing connection to any network, the device was known to be continuously probing for new networks. We also knew that the OUI of the device was 'Google' and the device was regularly randomising its MAC address, thus providing us an excellent reference with which we could optimise the parameters for our clustering algorithm. We then, by increasing the thresholds in steps of 1, found the minimum possible threshold for time and sequence numbers at which the algorithm clusters the reference device properly without over clustering the other probe requests. This process is shown in [Figure 3](#). We observed that the threshold for time  $\alpha$  and the threshold for sequence numbers,  $\beta$  are 16 s and 60, respectively. This was undertaken on top the filtering done based on signal strength, and only for the probe requests with randomised MAC addresses. [Figure 4](#) shows the results of this clustering process on a small set of randomised probe requests. The probe requests with different randomised MAC address are shown by the coloured points, and the lines joining them show that those probe requests were most likely be generated by the same device. We finally aggregated the probe requests as we did before but with the device signature rather than MAC addresses. This results in a footfall count with a MAPE of -18% compared to the manual count. A comparison of minute by minute counts resulting from different filtering processes along with the ground truth is shown in [Figure 4](#) illustrating the promising effectiveness of the methods.

To conclude, from the pilot study we found that both the filtering and the clustering methods we devised worked on complex real-world data and resulted in final pedestrian



**Figure 3.** The clustering process was repeated with both increasing sequence number threshold ( $\alpha$ ) and time threshold ( $\beta$ ) until we arrived at the lowest parameters where the know device (black line) is clustered as a single device.



**Figure 4.** The results of the pilot study demonstrating the validity of the methodology.

counts within a MAPE of 20%. We also found that ‘k-means’ and ‘quantile’ are the best algorithms for clustering signal strengths. Finally, we observed that the best thresholds for time and sequence numbers in the clustering algorithm are around 16 and 60, respectively.

## 5. Case study implementation

The methodology set out above was implemented in five different Central London locations at different times. Sensors were installed and data collected for extended periods of time. We also carried out manual counting at these locations across different times of the day. We then applied the methodologies discussed earlier to arrive at estimated pedestrian footfall and compared them with the corresponding manual counts. We finally evaluated the effectiveness of the processes with the Mean Absolute Percentage Error (MAPE) at the locations and report our findings below.

The locations at which the data were collected are shown in [Table 2](#). The locations were chosen for their diverse site conditions and unique sources of noise around the potential location of the sensors. The position of the sensor at these locations with respect to the context is shown [Figure 6](#). We can see that Location 5 is the ‘cleanest’ with one clear stationary source of noise (phone shop) while location 2 is the most complex due to the proximity of seating areas to the sensor. The sensors were operational throughout February and March, while manual counts were conducted in these locations in half-hour sessions on

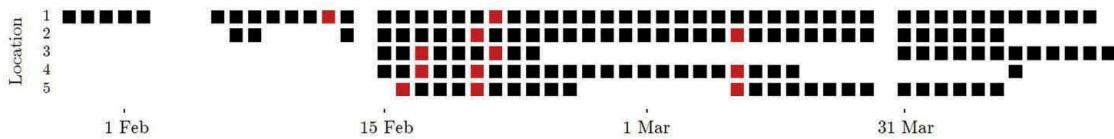
**Table 2.** Locations where sensors were installed, volume and speed of probe requests collected by the sender and total pedestrians manually counted. The data occupies around 1.8 GB on disk when encoded in text format.

ID	Location	Type	Installation notes	Probe Requests x10 <sup>6</sup> (per min)	Footfall No. (per min)
1	Camden High Street	Phone Shop	Bus stop in front	9.9 (297)	3683 (33)
2	Central St.Giles	Restaurant	Seating area on both sides	3.9 (169)	0346 (05)
3	Holborn Station	Info. Kiosk	Overlooks station entrance	4.3 (303)	2956 (46)
4	Brunswick Center	Fast Food	Has seating area on one side	3.4 (210)	0960 (12)
5	The Strand	Tea Shop	Has phone shop next door	8.4 (382)	1969 (21)

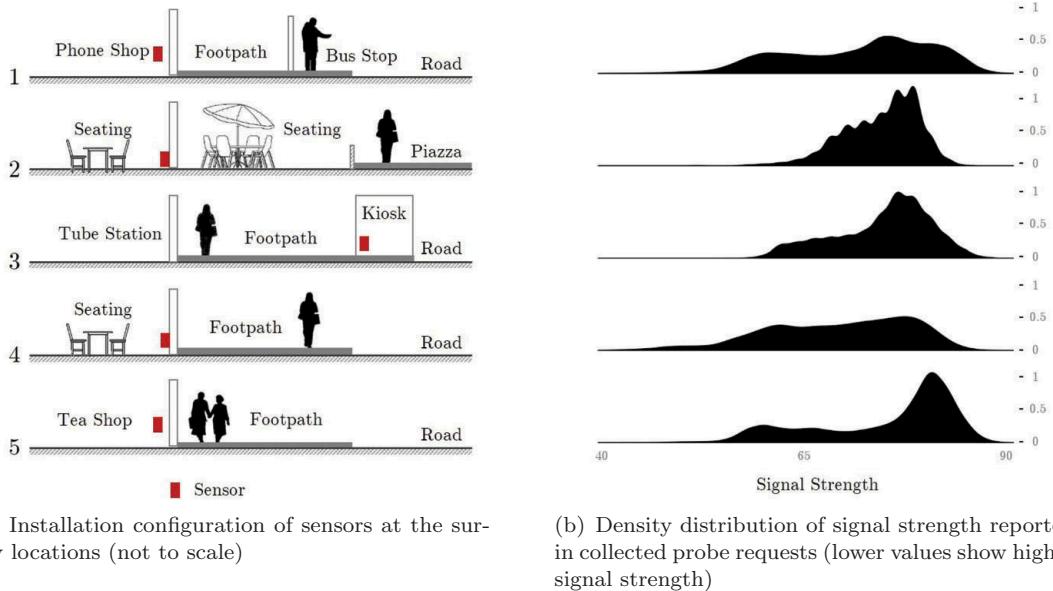
at least two different days. For the purposes of comparing with ground truth, we considered the data from sensors which correspond to the 12 sets of available manual counts. The schedule of data collection is shown in [Figure 5](#).

We begin by looking at the distribution of the signal strength reported by the probe requests across the locations. From the density plot shown in [Figure 6](#), we can observe that there is a clear relation between the distribution of the signal strength and the distance and complexity of the source of noise. We can see that while location 5 shows a clean difference between low and high signal strengths, location 2 is almost normally distributed. Intuitively we expected that location 2 and 4 would be harder to classify than locations 1, 3 and 5. We ran the k-means clustering algorithm and filtered out the probe requests which were randomised and had signal strengths less than the second break (threshold). It is important to note that we were dealing with relative thresholds of signal strengths which can vary with location and time of the analysis. We then aggregated the probe requests by counting the number of Unique MAC addresses present in every minute. We also removed devices that dwelled around the sensor by removing the MAC addresses which reappeared within the previous hour.

The results of the first stage of the filtering process along with the thresholds are shown in [Table 3](#). Confirming our intuition, we see that location 2 has the most MAPE followed by location 4, while the rest of them have highly reduced MAPE. It is significant that this method alone reduces our margin of error by 50–100% from the raw counts without any cleaning. This



**Figure 5.** Data collection schedule showing the days when sensors were active at their corresponding locations. The red squares show that manual counting of pedestrians was also done on that day.



**Figure 6.** Distribution of signal strengths across locations.

**Table 3.** Results of footfall estimation at each location as Mean Absolute Percentage Error (MAPE) after each step of the filtering process.

Sensor	Signal strength threshold (-dBm)	Adjustment factor	MAPE without any cleaning (%)	MAPE after filtering signal strength (%)	MAPE after filtering sequence numbers (%)	MAPE of final adjusted counts (%)
1	-70	1.25	259	22	-13	9
2	-74	0.51	928	396	206	55
3	-72	1.60	87	-19	-31	10
4	-70	0.88	498	142	52	33
5	-72	0.80	473	84	38	11

makes the signal strength filtering a quick and ideal method for practical applications, one which does not require absolute numbers such as creating large aggregated indexes to show long-term trends. We also found that the success of the signal strength filtering can be improved significantly by installing sensors so that the pedestrians and source of noise are at different distances from the sensor. This ensures that the distribution of signal strengths within the field of measurement is distinct from that of the surroundings.

We then ran the sequence numbers based clustering process on the rest of the probe requests to reduce the MAPE by almost 50–100% on all the sensors except for location 3. Location 3 is an outlier among all the other sensors since it is the only one with a large amount of pedestrians very close to the sensor. This may be the reason behind the over-filtering observed in the previous process. We finally ran the calibration process where we calculated the adjustment factors from the ratio between the manual counts to the sensor-based counts for the sample period as shown in **Table 3** and a comparison of the filtering process with the ground truth is illustrated in **Figure 7**. We used them to adjust the counts to achieve a MAPE ranging from 10% to 50%. We observed that the sensors with people moving right next to them tend to under-count with our methodology, while sensors with seating next to them tend to over-count significantly. However, using the filtering process, we can reduce the error to almost 10% closer to that of the ground truth.

## 6. Conclusion

Sentient technologies make measurement of the human activities that are the life blood of the smart city possible. Yet the data that they harvest are frequently relevant only to the sub-groups within society that avail themselves of particular goods and services – such as social media applications, transport modes or retail offers. In each of these cases, it is necessary to remember that the resulting data are by-products of consumer transactions, and will as a consequence, only pertain to users of the relevant goods or services. If the smart city is to be socially inclusive, it, therefore, follows that sentient data must represent entire populations, whether by design or by triangulation with external, population wide, sources. This is a non-trivial task, since the ebbs and flows of smart device-enabled citizens rarely pertain to any clearly defined population in either administrative or functional terms (Massam 1975).

Our objective here has been to collect, rather than re-use, data on smart city functioning, by recording Wi-Fi probes and ultimately reconciling them with manual counts in order to infer ambient populations. The internal validation methodology set out in the technical sections of this paper, allied to external validation from pedestrian counts, renders the method inclusive and robust when recording activity levels in retail centres in real time. We have described the collection and processing of a novel

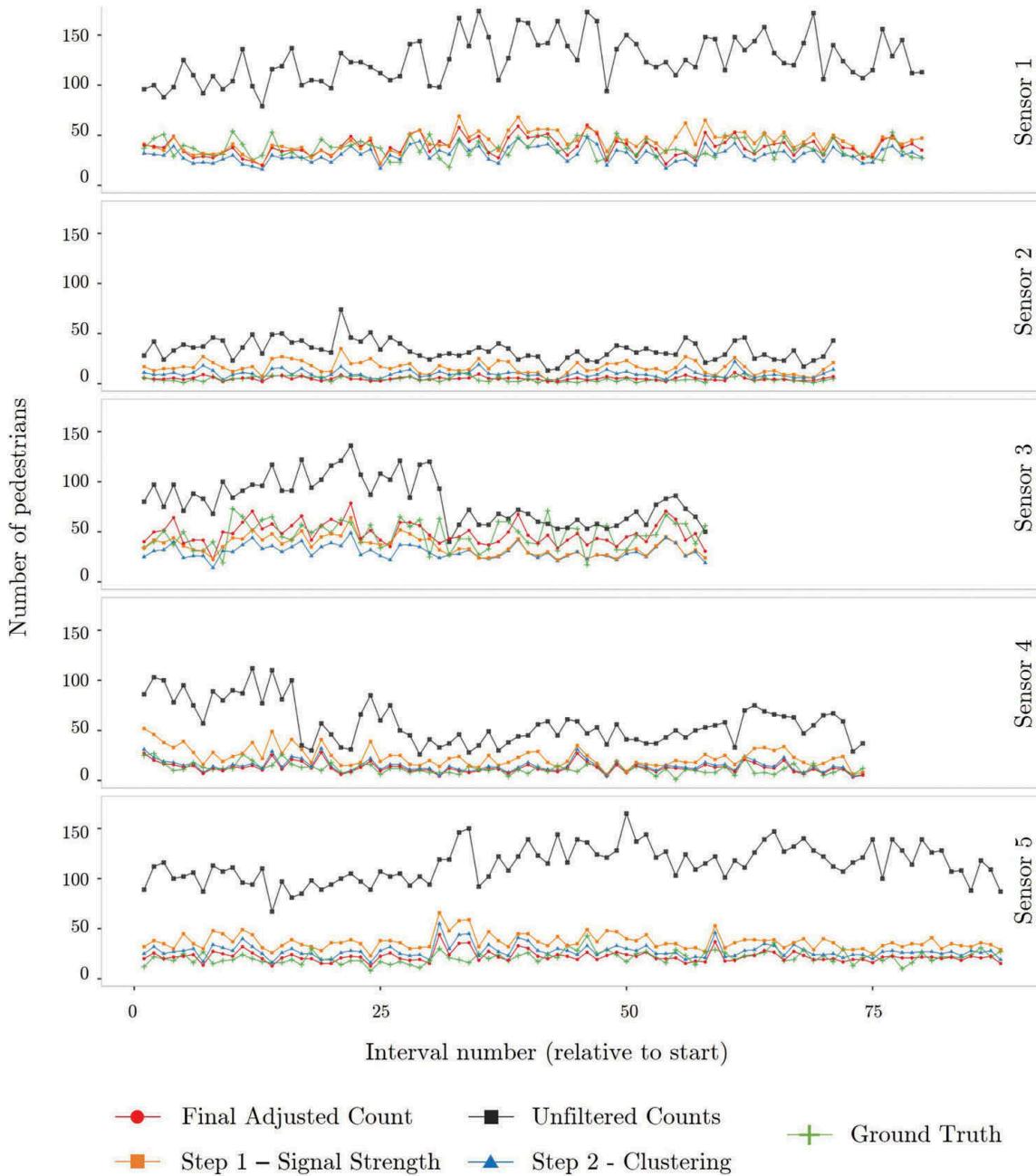
consumer Big Dataset that enables valid measures of levels of footfall activity which has been scaled across a wide network of sensors (Longley *et al.* 2018). In both conceptual and technical terms, it illustrates the ways in which passively collected consumer data can be ‘hardened’ to render them robust and reliable by using related procedures of internal and external validation.

Internal validation addresses the issues of screening out device probes that do not indicate footfall, and the further screening of device probes to ‘fingerprint’ the effects of MAC randomization. It is important to note that the filtering process work based solely on the information present in the probe requests and their temporal distribution. This ensures that although the mobile devices were uniquely identified, there was no further personal data generated by linking the probe requests to the users of the mobile devices. This method essentially gave us a way to estimate the footfall in real-time without identifying or tracking the mobile devices themselves. External validation then entailed reconciling adjusted counts with the footfall observed at sample locations. This procedure makes it possible to generalise from locations at which manual footfall surveys are conducted to all others in the system, and to develop a classification of device locations that are more or less susceptible to noise generation.

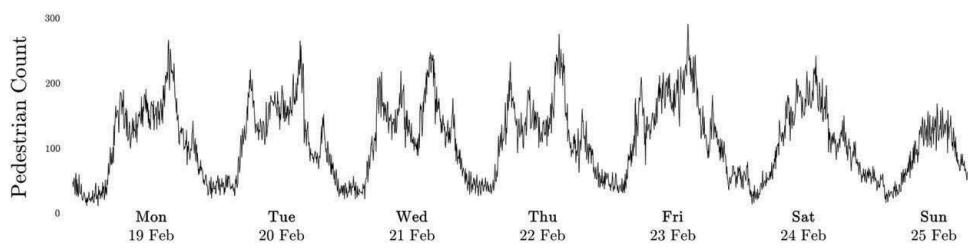
This Wi-Fi-based footfall counting methodology offers a large number of applications and benefits for real-time spatial analysis. Since Wi-Fi-based sensors are inexpensive and the data model is scalable, it is possible to use this methodology for a large network of sensors to gather granular data on pedestrian footfall. A snapshot showing a week’s worth of precise footfall in the area around Charring cross, London is shown in [Figure 8](#) in order to demonstrate the potential for such a dataset. Projects such as SmartStreetSensors (Longley *et al.* 2018), may utilise this methodology to overcome the challenges introduced by the implementation of MAC address randomisation.

The vicissitudes of MAC randomisation, and the provisions of privacy legislation such as EU General Data Protection Regulations mitigate against tracking individuals across the smart city using this approach. This can be modelled using agent-based methods (Heppenstall *et al.* 2011), however. In our own research, we have also begun to link store time-lagged till receipts to footfall, and have used such data to better understand the dwell times that characterise such different retail uses as stores with window displays and fast food restaurants. Such analysis not only provides a more nuanced picture of movement through retail areas but also enables valorisation of microsites within retail centres. In the UK, for example, this is of immediate practical importance in evaluating business rates on properties and has still wider implications for the setting of retail unit rental values. There are obvious extensions to understanding the ebbs and flows of activities in the 24-h smart city such as understanding urban mobility (Gariazzo *et al.* 2019) and conceptualising them with a people dimension (Nam and Pardo 2011).

More broadly still, extensions to this strand of smart city research are likely to seek to differentiate the quality of different elements within footfall according to mission, e.g., travel to adjacent workplace zones, and leisure, and personal characteristics such as spending power. In this respect, future research may not only simulate linkage of harmonised footfall counts between sensor locations, but also link these, in turn, to disaggregate origin-destination matrices for bikeshare and other public transport modes. Our own investigations will consider these and other challenges to understanding the functioning of the sentient city.



**Figure 7.** Comparison of the filtering process with the ground truth in all the locations.



**Figure 8.** A week of pedestrian footfall at the Strand, London collected by the methodology. The counts are aggregated for 5-minute intervals.

# Bibliography

Command-line tools can be 235x faster than your hadoop cluster, Jan 2014.  
URL <https://bit.ly/2s2XZYI>.

Michael Abbott-Jard, Harpal Shah, and Ashish Bhaskar. Empirical evaluation of bluetooth and wifi scanning for road transport. In *Australasian Transport Research Forum (ATRF), 36th, 2013, Brisbane, Queensland, Australia*, page 14, 2013a.

Michael Abbott-Jard, Harpal Shah, and Ashish Bhaskar. Empirical evaluation of bluetooth and Wi-Fi scanning for road transport. In *Australasian Transport Research Forum (ATRF), 36th, 2013, Brisbane, Queensland, Australia*, page 14, 2013b.

Naeim Abedi, Ashish Bhaskar, and Edward Chung. Bluetooth and wi-fi mac address based crowd data collection and monitoring : Benefits , challenges and enhancement. *Australasian Transport Research Forum 2013 Proceedings 2*, (October):1–17, 2013. URL <http://www.patrec.org/atrf.aspx>.

John M Abowd, John Haltiwanger, and Julia Lane. Integrated longitudinal employer-employee data for the united states. *The American Economic Review*, 94(2):224–229, 2004. ISSN 0002-8282.

Florian Adamsky, Tatiana Retunskaya, Stefan Schiffner, Christian Köbel, and Thomas Engel. Wlan device fingerprinting using channel state information (csi). In *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, pages 277–278. ACM, 2018.

Rein Ahas and Ülar Mark. Location based services - new challenges for planning and public administration? *Futures*, 37(6):547–561, 2005. ISSN 00163287. DOI: [10.1016/j.futures.2004.10.012](https://doi.org/10.1016/j.futures.2004.10.012).

Mohammed Alahmadi, Peter Atkinson, and David Martin. Estimating the spatial distribution of the population of riyadh, saudi arabia using remotely sensed built land cover and height data. *Computers, Environment and Urban Systems*, 41:167–176, 2013. ISSN 0198-9715.

David E Alexander. *Principles of emergency planning and management*. Oxford University Press on Demand, 2002. ISBN 0195218388.

Alexander Amini, Kevin Kung, Chaogui Kang, Stanislav Sobolevsky, and Carlo Ratti. The impact of social segregation on human mobility in developing and industrialized regions. *EPJ Data Science*, 3(1):1, 2014. ISSN 2193-1127.

Aleksandar Antonic, Ivana Podnar Zarko, and Domagoj Jakobovic. Inferring presence status on smartphones: The big data perspective. In *Proceedings - International Symposium on Computers and Communications*, pages 600–605, 2013. ISBN 9781479937554. DOI: 10.1109/ISCC.2013.6755013.

Irina Arhipova, Gundars Berzins, Edgars Brekis, Martins Opmanis, Juris Binde, Jevgenija Kravcova, and Inna Steinbuka. Pattern identification by factor analysis for regions with similar economic activity based on mobile communication data. In *Future of Information and Communication Conference*, pages 561–569. Springer, 2018.

Kai O Arras, Slawomir Grzonka, Matthias Luber, and Wolfram Burgard. Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 1710–1715. IEEE, 2008. ISBN 1424416469.

Dani Arribas-Bel and Emmanouil Tranos. New approaches to measure the spatial structure (s) of cities. 2015.

Daniel Arribas-Bel. Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*, 49:45–53, 2014. ISSN 01436228. DOI: 10.1016/j.apgeog.2013.09.012.

Daniel Arribas-Bel and Fernando Sanz-Gracia. The validity of the monocentric city model in a polycentric age: Us metropolitan areas in 1990, 2000 and 2010. *Urban Geography*, 35(7):980–997, 2014. ISSN 0272-3638.

Yasuo Asakura and Eiji Hato. Tracking survey for individual travel behaviour using mobile communication instruments. *Transportation Research Part C: Emerging Technologies*, 12(3):273–291, 2004. ISSN 0968-090X.

Fereshteh Asgari, Vincent Gauthier, and Monique Becker. A survey on human mobility and its applications. *arXiv preprint arXiv:1307.0814*, 2013.

Paramvir Bahl and Venkata N Padmanabhan. Radar: An in-building rf-based user location and tracking system. In *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 2, pages 775–784. Ieee, 2000. ISBN 0780358805.

Udana Bandara, Mikio Hasegawa, Masugi Inoue, Hiroyuki Morikawa, and Tomonori Aoyama. Design and implementation of a bluetooth signal strength based location sensing system. In *Radio and Wireless Conference, 2004 IEEE*, pages 319–322. IEEE, 2004. ISBN 0780384512.

Tengfei Bao, Huanhuan Cao, Qiang Yang, Enhong Chen, and Jilei Tian. Mining significant places from cell id trajectories: A geo-grid based approach. In *2012 IEEE 13th International Conference on Mobile Data Management*, pages 288–293. IEEE, 2012. ISBN 1467317969.

Marco V Barbera, Alessandro Epasto, Alessandro Mei, Vasile C Perta, and Julinda Stefa. Signals from the crowd: uncovering social relationships through smartphone probes. In *Proceedings of the 2013 conference on Internet measurement conference*, pages 265–276. ACM, 2013. ISBN 145031953X.

- Michael Batty. Invisible cities. *Environment and Planning B: Planning and Design*, 17(2):127–130, 1990.
- Michael Batty. Virtual geography. *Futures*, 29(4-5):337–352, 1997. ISSN 00163287. DOI: 10.1016/S0016-3287(97)00018-9. URL <http://www.sciencedirect.com/science/article/pii/S0016328797000189>.
- Michael Batty. The pulse of the city. *Environment and Planning B: Planning and Design*, 37(4):575–577, 2010. ISSN 0265-8135.
- Michael Batty. *The new science of cities*. Mit Press, 2013a. ISBN 0262019523.
- Michael Batty. The future cities agenda. *Environ. Plann. B Plann. Des*, 40: 191–194, 2013b.
- Michael Batty, Kay W Axhausen, Fosca Giannotti, Alexei Pozdnoukhov, Armando Bazzani, Monica Wachowicz, Georgios Ouzounis, and Yuval Portugali. Smart cities of the future. *The European Physical Journal Special Topics*, 214(1):481–518, 2012. ISSN 1951-6355.
- Anil Bawa-Cavia. Sensing the urban: using location-based social network data in urban analysis. In *Workshop on Pervasive and Urban Applications (PURBA)*, pages 1–7, 2011.
- Johannes K Becker, David Li, and David Starobinski. Tracking anonymized bluetooth devices. *Proceedings on Privacy Enhancing Technologies*, 1:17, 2019.
- Richard A Becker, Ramon Caceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. Route classification using cellular handoff patterns. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 123–132. ACM, 2011a. ISBN 1450306306.
- Richard A Becker, Ramon Caceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4):18–26, 2011b. ISSN 1536-1268. DOI: 10.1109/MPRV.2011.44. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5928310>.
- Alan Bensky. Wireless positioning technologies and applications, artech house. Inc., Norwood, MA, 2007.
- Luca Bertolini and Martin Dijst. Mobility environments and network cities. *Journal of urban design*, 8(1):27–43, 2003.
- Luís M A Bettencourt. The origins of scaling in cities. *science*, 340(6139): 1438–1441, 2013. ISSN 0036-8075.
- Luis M A Luís M.A. Bettencourt. The uses of big data in cities. *Big Data*, 2(1):12–22, 2014. ISSN 2167-6461. DOI: 10.1089/big.2013.0042. URL <http://online.liebertpub.com/doi/abs/10.1089/big.2013.0042>.
- Budhendra Bhaduri, Edward Bright, Phillip Coleman, and Jerome Dobson. Landscan. *Geoinformatics*, 5(2):34–37, 2002.
- Budhendra Bhaduri, Eddie Bright, and Phil Coleman. Development of a high resolution population dynamics model. *Geocomputation 2005*, 2005.

- Budhendra Bhaduri, Edward Bright, Phillip Coleman, and Marie L. Urban. Landscan usa: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics. In *GeoJournal*, volume 69, pages 103–117. Springer, 2007. ISBN 0165-0009. DOI: [10.1007/s10708-007-9105-9](https://doi.org/10.1007/s10708-007-9105-9).
- Bastian Bloessl, Christoph Sommer, Falko Dressier, and David Eckhoff. The scrambler attack: A robust physical layer attack on location privacy in vehicular networks. In *Computing, Networking and Communications (ICNC), 2015 International Conference on*, pages 395–400. IEEE, 2015. ISBN 1479969591.
- R Bolla, F Davoli, and F Giordano. Estimating road traffic parameters from mobile communications. In *Proceedings 7th World Congress on ITS, Turin, Italy*, 2000.
- Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011. URL <http://vis.stanford.edu/papers/d3>.
- Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006. ISSN 0028-0836.
- Caroline O Buckee, Andrew J Tatem, Justin Lessler, Ottar N Bjornstad, Bryan T Grenfell, Janeth Kombich, Nathan Eagle, C J E Metcalf, and Amy Wesolowski. Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. *Proceedings of the National Academy of Sciences*, 112(35):11114–9, 2015. ISSN 0027-8424. DOI: [10.1073/pnas.1423542112](https://doi.org/10.1073/pnas.1423542112). URL <http://doi.org/10.1073/pnas.1423542112>.
- Nirupama Bulusu, John Heidemann, and Deborah Estrin. Gps-less low-cost outdoor localization for very small devices. *IEEE personal communications*, 7(5):28–34, 2000. ISSN 1070-9916.
- N Caceres, J P Wideberg, and F G Benitez. Deriving origin destination data from a mobile phone network. *IET Intelligent Transport Systems*, 1(1):15–26, 2007. ISSN 1751-956X.
- Quin Cai and Jake K Aggarwal. Tracking human motion using multiple cameras. In *International Conference on Pattern Recognition*, volume 13, pages 68–72. Citeseer, 1996. ISBN 1051-4651.
- Francesco Calabrese, Francisco C Pereira, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. The geography of taste: analyzing cell-phone mobility and social events. In *International Conference on Pervasive Computing*, pages 22–37. Springer, 2010.
- Francesco Calabrese, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):36–44, 2011. ISSN 15361268. DOI: [10.1109/MPRV.2011.41](https://doi.org/10.1109/MPRV.2011.41).
- Francesco Calabrese, Mi Diao, Giusy Di Lorenzo, Jr. Ferreira Joseph, and Carlo Ratti. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, 26(0):301–313, 2013.

Francesco Calabrese, Laura Ferrari, and Vincent D. Blondel. Urban sensing using mobile phone network data: a survey of research. *ACM Computing Surveys (CSUR)*, 47(2):25, 2015. ISSN 0360-0300. DOI: 10.1145/2655691. URL <http://dl.acm.org/citation.cfm?doid=2658850.2655691>.

Andrew T Campbell, Shane B Eisenman, Kristff Fodor, Nicholas D Lane, Hong Lu, Emiliano Miluzzo, Mirco Musolesi, Ronald A Peterson, and Xiao Zheng. Transforming the social networking experience with sensing presence from mobile phones. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, pages 367–368. ACM, 2008.

Julián Candia, Marta C. González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. Uncovering individual and collective human dynamics from mobile phone records. *J. Phys. A: Math. Theor.*, 41(22):224015–11, 2008. ISSN 1751-8113. DOI: 10.1088/1751-8113/41/22/224015.

M Castells. Grassrooting the space of flows. i wheeler, aoyama and warf [eds.] cities in the telecommunications age, 2000.

M Castells. *The Rise of the Network Society*, volume I. 2010. ISBN 9781405196864. DOI: 10.2307/1252090. URL <http://www.lavoisier.fr/livre/notice.asp?depuis=e.lavoisier.fr&id=9781405196864>.

CDRC. Smart street sensor project. [goo.gl/E4tR8o](http://goo.gl/E4tR8o), 2016. [Online; accessed 31-January-2018].

Junghoon Chae, Dennis Thom, Harald Bosch, Yun Jang, Ross Maciejewski, David S Ebert, and Thomas Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 143–152. IEEE, 2012. ISBN 1467347523.

Jie Chen, Tao Pei, Shih-Lung Shaw, Feng Lu, Mingxiao Li, Shifen Cheng, Xiliang Liu, and Hengcai Zhang. Fine-grained prediction of urban population using mobile phone location data. *International Journal of Geographical Information Science*, 32(9):1770–1786, 2018.

K Chen. An approach to linking remotely sensed data and areal census data. *International Journal of Remote Sensing*, 23(1):37–48, 2002. ISSN 0143-1161.

Linsong Cheng and Jiliang Wang. How can i guard my ap?: non-intrusive user identification for mobile devices using wifi signals. In *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pages 91–100. ACM, 2016.

Ningning Cheng, Prasant Mohapatra, Mathieu Cunche, Mohamed Ali Kaafar, Roksana Boreli, and Srikanth Krishnamurthy. Inferring user relationship from hidden information in wlans. In *Military Communications Conference, 2012-MILCOM 2012*, pages 1–6. IEEE, 2012. ISBN 1467317314.

Peng Cheng, Zhijun Qiu, and Bin Ran. Particle filter based traffic state estimation using cell phone network data. In *2006 IEEE Intelligent Transportation Systems Conference*, pages 1047–1052. IEEE, 2006. ISBN 1424400937.

Guanghua Chi, Jean-Claude Thill, Daoqin Tong, Li Shi, and Yu Liu. Uncovering regional characteristics from mobile phone data: A network science approach. *Papers in Regional Science*, 2014. ISSN 1435-5957.

Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

Tom Chothia and Vitaliy Smirnov. A traceability attack against e-passports. In *International Conference on Financial Cryptography and Data Security*, pages 20–34. Springer, 2010. ISBN 3642145760.

Cisco. *Wi-Fi Location-Based Services 4.1 Design Guide*. Cisco Systems Inc., 2008. URL <https://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Mobility/WiFiLBS-DG.html>.

Gerald Combs and Contributors. Wireshark - network protocol analyzer. <https://www.wireshark.org/about.html>, 2018.

Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

Tomas Crols and Nick Malleson. Quantifying the ambient population using hourly population footfall data and an agent-based model of daily mobility. *Geoinformatica*, pages 1–20, 2019.

Mathieu Cunche. I know your mac address: targeted tracking of individual using wi-fi, 2014. ISSN 22638733.

Mathieu Cunche, Mohamed Ali Kaafar, and Roksana Boreli. I know who you will meet this evening! linking wireless devices using wi-fi probe requests. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2012 IEEE International Symposium on*, pages 1–9. IEEE, 2012. ISBN 1467312398.

Mathieu Cunche, Mohamed-Ali Kaafar, and Roksana Boreli. Linking wireless devices using information contained in wi-fi probe requests. *Pervasive and Mobile Computing*, 11:56–69, 2014. ISSN 1574-1192.

S Cutter, Douglas B Richardson, and Thomas J Wilbanks. The changing landscape of fear. *The geographical dimensions of terrorism*, pages 1–5, 2006.

Peng Dai, Yuan Yang, Manyi Wang, and Ruqiang Yan. Combination of dnn and improved knn for indoor location fingerprinting. *Wireless Communications and Mobile Computing*, 2019, 2019.

Deloitte. Mobile consumer survey - united kingdom, 2018. URL <https://www.deloitte.co.uk/mobileuk/>.

Levent Demir, Mathieu Cunche, and Cédric Lauradoux. Analysing the privacy policies of Wi-Fi trackers. In *Proceedings of the 2014 workshop on physical analytics*, pages 39–44. ACM, 2014.

Merkebe Getachew Demissie, Gonçalo Homem de Almeida Correia, and Carlos Bento. Intelligent road traffic status detection system through cellular networks handover information: An exploratory study. *Transportation research part C: emerging technologies*, 32:76–88, 2013. ISSN 0968-090X.

Zhihong Deng, Yun Cao, Pengyu Wang, and Bo Wang. An improved heuristic drift elimination method for indoor pedestrian positioning. *Sensors*, 18(6):1874, 2018.

Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014. ISSN 0027-8424. DOI: 10.1073/pnas.1408439111.

Adriano Di Luzio, Alessandro Mei, and Julinda Stefa. Mind your probes: De-anonymization of large crowds through smartphone wifi probe requests. In *Proceedings - IEEE INFOCOM*, volume 2016-July, 2016. ISBN 9781467399531. DOI: 10.1109/INFOCOM.2016.7524459.

Bram Dil and Paul J M Havinga. Stochastic radio interferometric positioning in the 2.4 ghz range. In *SenSys*, pages 108–120, 2011.

Vaibhav Dinesh, Prakash Tripathi, and Ajit Singh. Device-location estimation based on rssi measurements over wifi and bluetooth, March 2 2017. US Patent App. 15/056,902.

Trinh Minh Tri Do and Daniel Gatica-Perez. The places of our lives: Visiting patterns and automatic labeling from longitudinal smartphone data. *IEEE Transactions on Mobile Computing*, 13(3):638–648, 2014. ISSN 15361233. DOI: 10.1109/TMC.2013.19.

Jerome E Dobson and Peter F Fisher. Geoslavery. *IEEE Technology and Society Magazine*, 22(1):47–52, 2003. ISSN 0278-0097.

Jerome E Dobson, Edward A Bright, Phillip R Coleman, Richard C Durfee, and Brian A Worley. Landscan: a global population database for estimating populations at risk. *Photogrammetric engineering and remote sensing*, 66(7):849–857, 2000. ISSN 0099-1112.

Jerome E Dobson, Eddie A Bright, Phil R Coleman, and Budhendra L Bhaduri. Landscan2000: A new global population geography. Technical report, 2003.

Matt Duckham and Lars Kulik. Location privacy and location-aware computing. *Dynamic & mobile GIS: investigating change in space and time*, 3:35–51, 2006.

Nathan Eagle and Alex Sandy Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009. ISSN 03405443. DOI: 10.1007/s00265-009-0739-0.

Nature Editorial. A flood of hard data. *Nature*, 435:698, 2008. ISSN 0028-0836. DOI: 10.1038/453698a.

A Elgohary. On detecting device-free entities using wifi signals. *ece.uwaterloo.ca*, 2013. URL <https://ece.uwaterloo.ca/~aelgohar/stat841-report.pdf>. ISBN 5Cnpapers3://publication/uuid/D6821814-0041-47E6-9A26-96A32F41B07F.

Gorkem Erinc, Benjamin Balaguer, and Stefano Carpin. Heterogeneous map merging using wifi signals. In *IEEE International Conference on Intelligent Robots and Systems*, pages 5258–5264, 2013. ISBN 9781467363587. DOI: 10.1109/IROS.2013.6697117.

Sara Irina Fabrikant. Towards an understanding of geovisualization with dynamic displays: Issues and prospects. In *AAAI Spring Symposium: Reasoning with Mental and External Diagrams: Computational Modeling and Spatial Assistance*, pages 6–11, 2005.

Katayoun Farrahi and Daniel Gatica-Perez. Probabilistic mining of socio-geographic routines from mobile phone data. *IEEE Journal of Selected Topics in Signal Processing*, 4(4):746–755, 2010.

Emilio Ferrara, Pasquale De Meo, Salvatore Catanese, and Giacomo Fiumara. Visualizing criminal networks reconstructed from mobile phone records. In *CEUR Workshop Proceedings*, volume 1210, 2014.

Donald L Foley. Urban daytime population: a field for demographic-ecological analysis. *Social Forces*, pages 323–330, 1954. ISSN 0037-7732.

Jason Franklin, Damon McCoy, Parisa Tabriz, Vicentiu Neagoe, Jamie V Randwyk, and Douglas Sicker. Passive data link layer 802.11 wireless device driver fingerprinting. In *USENIX Security Symposium*, volume 3, pages 16–89, 2006.

Julien Freudiger. How talkative is your mobile device?: An experimental study of Wi-Fi probe requests. In *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, WiSec ’15, pages 8:1–8:6, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3623-9. DOI: 10.1145/2766498.2766517. URL <http://doi.acm.org/10.1145/2766498.2766517>.

Yuki Fukuzaki, Masahiro Mochizuki, Kazuya Murao, and Nobuhiko Nishio. A pedestrian flow analysis system using wi-fi packet sensors to a real environment. *Proceedings of UbiComp-Adjunct*, pages 721–730, 2014. DOI: 10.1145/2638728.2641312. URL [http://ubicomp.org/ubicomp2014/proceedings/ubicomp\\_{\\\_}adjunct/workshops/HASCA/p721-fukuzaki.pdf](http://ubicomp.org/ubicomp2014/proceedings/ubicomp_{\_}adjunct/workshops/HASCA/p721-fukuzaki.pdf).

Yuuaki Fukuzaki, Masahiro Mochizuki, Kazuya Murao, and Nobuhiko Nishio. Statistical analysis of actual number of pedestrians for wi-fi packet-based pedestrian flow sensing. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pages 1519–1526, 2015. ISBN 978-1-4503-3575-1. DOI: 10.1145/2800835.2801623. URL <http://doi.acm.org/10.1145/2800835.2801623>.

Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144, 2015.

Claudio Gariazzo, Armando Pelliccioni, and Maria Paola Bogliolo. Spatiotemporal analysis of urban mobility using aggregate mobile phone derived presence and demographic data: A case study in the city of rome, italy. *Data*, 4(1):8, 2019.

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *Proceedings of the 19th ACM Symposium on Operating Systems Principles*, pages 20–43, Bolton Landing, NY, 2003.

Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 330–339. ACM, 2007. ISBN 1595936092.

Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Fabio Pinelli, Chiara Renso, Salvatore Rinzivillo, and Roberto Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal—The International Journal on Very Large Data Bases*, 20(5):695–719, 2011. ISSN 1066-8888.

Fabien Girardin, Francesco Calabrese, Filippo Dal Fiore, Carlo Ratti, and Josep Blat. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive computing*, 7(4):36–43, 2008. ISSN 1536-1268.

Fabien Girardin, Andrea Vaccari, Alexandre Gerber, Assaf Biderman, and Carlo Ratti. Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate. In *Intl. Conference on Computers in Urban Planning and Urban Management*, 2009.

Genevieve Giuliano and Kenneth A. Small. Subcenters in the los angeles region. *Regional Science and Urban Economics*, 21(2):163–182, 1991. ISSN 01660462. DOI: 10.1016/0166-0462(91)90032-I.

Theodore S Glickman. A methodology for estimating time-of-day variations in the size of a population exposed to risk. *Risk Analysis*, 6(3):317–324, 1986. ISSN 1539-6924.

Hongmian Gong, Cynthia Chen, Evan Bialostozky, and Catherine T Lawson. A gps/gis method for travel mode detection in new york city. *Computers, Environment and Urban Systems*, 36(2):131–139, 2012. ISSN 0198-9715.

Marta C. González, Cesar A César A. Hidalgo, Albert-László Barabási, Marta C Gonzalez, Cesar A César A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008. ISSN 0028-0836. DOI: 10.1038/nature06958. URL <http://www.nature.com/nature/journal/v453/n7196/full/nature06958.html>.

Stephen Graham. Cities in the real-time age: The paradigm challenge of telecommunications to the conception and planning of urban space. *Environment and Planning A*, 29(1):105–127, 1997. ISSN 0308518X. DOI: 10.1068/a290105.

Stephen Graham and Patsy Healey. Relational concepts of space and place : Issues for planning theory and practice. *European Planning Studies*, 7(5): 623–646, 1999. ISSN 0965-4313. DOI: 10.1080/09654319908720542. URL <http://www.tandfonline.com/doi/abs/10.1080/09654319908720542>.

Stephen Graham and Simon Marvin. *Splintering Urbanism*. 2001. ISBN 0203452208. DOI: 10.4324/9780203452202. URL <http://books.google.com/books?id=6IdAAY9xqlgC&pgis=1>.

Steve Graham and Simon Marvin. *Telecommunications and the city: Electronic spaces, urban places*. Routledge, 2002. ISBN 1134813937.

Sebastian Grauwin, Stanislav Sobolevsky, Simon Moritz, István Góðor, and Carlo Ratti. Towards a comparative science of cities: Using mobile traffic records in new york, london, and hong kong. In *Computational Approaches for Urban Environments*, pages 363–387. 2015. ISBN 9783319114699. DOI: 10.1007/978-3-319-11469-9\_15.

Ben Greenstein, Damon McCoy, Jeffrey Pang, Tadayoshi Kohno, Srinivasan Seshan, and David Wetherall. Improving wireless privacy with an identifier-free link layer protocol. In *Proceedings of the 6th international conference on Mobile systems, applications, and services*, pages 40–53. ACM, 2008. ISBN 1605581399.

Marco Gruteser and Dirk Grunwald. Enhancing location privacy in wireless lan through disposable interface identifiers: a quantitative analysis. *Mobile Networks and Applications*, 10(3):315–325, 2005. ISSN 1383-469X.

Theo Haerder and Andreas Reuter. Principles of transaction-oriented database recovery. *ACM Comput. Surv.*, 15(4):287–317, December 1983. ISSN 0360-0300. DOI: 10.1145/289.291. URL <http://doi.acm.org/10.1145/289.291>.

Ali Haghani, Masoud Hamed, Kaveh Sadabadi, Stanley Young, and Philip Tarnoff. Data collection of freeway travel time ground truth with bluetooth sensors. *Transportation Research Record: Journal of the Transportation Research Board*, (2160):60–68, 2010. ISSN 0361-1981.

Jan Erik Håkegård, Tor Andre Myrvoll, and Tor Rune Skoglund. Statistical modelling for estimation of od matrices for public transport using wi-fi and apc data. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 1005–1010. IEEE, 2018.

Elaine J Hallisey. Cartographic visualization: an assessment and epistemological review\*. *The Professional Geographer*, 57(3):350–364, 2005. ISSN 0033-0124.

Mark Harrower. The cognitive limits of animated maps. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 42(4): 349–357, 2007. ISSN 0317-7173.

Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271, 2014. ISSN 1523-0406. DOI: 10.1080/15230406.2014.890072. URL <http://www.tandfonline.com/doi/full/10.1080/15230406.2014.890072?src=recsys>.

Kevin Hawley and Harold Moellering. A comparative analysis of areal interpolation methods. *Cartography and Geographic Information Science*, 32(4):411–423, 2005. ISSN 1523-0406.

S I Hay, A M Noor, A Nelson, and A J Tatem. The accuracy of human population maps for public health application. *Tropical Medicine & International Health*, 10(10):1073–1086, 2005. ISSN 1365-3156.

Tian He, Chengdu Huang, Brian M Blum, John A Stankovic, and Tarek Abdelzaher. Range-free localization schemes for large scale sensor networks. In *Proceedings of the 9th annual international conference on Mobile computing and networking*, pages 81–95. ACM, 2003. ISBN 1581137532.

Janne Heikkilä and Olli Silvén. A real-time system for monitoring of cyclists and pedestrians. *Image and Vision Computing*, 22(7):563–570, 2004. ISSN 0262-8856.

Alison J Heppenstall, Andrew T Crooks, Linda M See, and Michael Batty. *Agent-based models of geographical systems*. Springer Science & Business Media, 2011.

Cesar A Hidalgo and C Rodriguez-Sickert. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387(12):3017–3024, 2008. ISSN 0378-4371.

Hande Hong, Girisha Durrel De Silva, and Mun Choon Chan. Crowdprobe: Non-invasive crowd monitoring with Wi-Fi probe. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):115, 2018.

Sahar Hoteit, Stefano Secci, Stanislav Sobolevsky, Guy Pujolle, and Carlo Ratti. Estimating real human trajectories through mobile phone data. In *2013 IEEE 14th International Conference on Mobile Data Management*, volume 2, pages 148–153. IEEE, 2013. ISBN 1467360686.

IEEE. IEEE standard for information technology-telecommunications and information exchange between systems local and metropolitan area networks-specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. *IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)*, pages 1–3534, Dec 2016.

Md Shahadat Iqbal, Charisma F. Choudhury, Pu Wang, and Marta C. González. Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74, 2014. ISSN 0968090X. DOI: 10.1016/j.trc.2014.01.002.

Chris Jacobs-Crisioni, Piet Rietveld, Eric Koomen, and Emmanouil Tranos. Evaluating the impact of land-use density and mix on spatiotemporal urban activity patterns: An exploratory study using mobile phone data.

*Environment and Planning A*, 46(11):2769–2785, 2014. ISSN 14723409. DOI: 10.1068/a130309p.

HV Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94, 2014.

Andreas Janecek, Karin A Hummel, Danilo Valerio, Fabio Ricciato, and Helmut Hlavacs. Cellular data meet vehicular traffic theory: location area updates and cell transitions for travel time estimation. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 361–370. ACM, 2012. ISBN 1450312241.

Davy Janssens. *Data Science and Simulation in Transportation Research*. IGI Global, 2013. ISBN 1466649216.

Bing Jia, Baoqi Huang, Hepeng Gao, Wuyungerile Li, and Lifei Hao. Selecting critical wifi aps for indoor localization based on a theoretical error analysis. *IEEE Access*, 2019.

Chen Jia, Yunyan Du, Siying Wang, Tianyang Bai, and Teng Fei. Measuring the vibrancy of urban neighborhoods using mobile phone data with an improved pagerank algorithm. *Transactions in GIS*, 11 2018. DOI: 10.1111/tgis.12515.

Peng Jia, Youliang Qiu, and Andrea E. Gaughan. A fine-scale spatial population distribution on the high-resolution gridded population surface and application in alachua county, florida. *Applied Geography*, 50:99–107, 2014. ISSN 01436228. DOI: 10.1016/j.apgeog.2014.02.009.

Bin Jiang and Xiaobai Yao. Location-based services and gis in perspective. *Computers, Environment and Urban Systems*, 30(6):712–725, 2006. ISSN 0198-9715.

Shan Jiang, Joseph Ferreira, and Marta C. González. Clustering daily patterns of human activities in the city. In *Data Mining and Knowledge Discovery*, volume 25, pages 478–510. Springer, 2012. ISBN 1384-5810. DOI: 10.1007/s10618-012-0264-z.

Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Fazzoli, and Marta C González. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing*, page 2. ACM, 2013. ISBN 1450323316.

Shan Jiang, Ana Alves, Filipe Rodrigues, Joseph Ferreira, and Francisco C Pereira. Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*, 53:36–46, 2015. ISSN 0198-9715.

Wang Jing, Wang Dianhai, Song Xianmin, and Sun Di. Dynamic od expansion method based on mobile phone location. In *Intelligent Computation Technology and Automation (ICICTA), 2011 International Conference on*, volume 1, pages 788–791. IEEE, 2011. ISBN 1612842895.

Warren C. Jochem, Kelly Sims, Edward A. Bright, Marie L. Urban, Amy N. Rose, Phillip R. Coleman, and Budhendra L. Bhaduri. Estimating traveler populations at airport and cruise terminals for population distribution and dynamics. *Natural Hazards*, 68(3):1325–1342, 2013. ISSN 0921030X. DOI: 10.1007/s11069-012-0441-9.

Chaogui Kang, Xiujun Ma, Daoqin Tong, and Yu Liu. Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and its Applications*, 391(4):1702–1717, 2012. ISSN 0378-4371.

Nobuo Kawaguchi. Wifi location information system for both indoors and outdoors. In *International Work-Conference on Artificial Neural Networks*, pages 638–645. Springer, 2009.

E.-K Kim and Am MacEachren. An index for characterizing spatial bursts of movements: A case study with geo-located twitter data. *Sites.Utexas.Edu*, (1), 2014. URL [http://sites.utexas.edu/amd2014/files/2014/02/2014\\_{\\_}GIScienceWorkshop{\\_}final{\\_}ekamm.pdf](http://sites.utexas.edu/amd2014/files/2014/02/2014_{_}GIScienceWorkshop{_}final{_}ekamm.pdf).

Minkyong Kim, David Kotz, Songkuk Kim, Kim Minkyong, David Kotz, and Kim Songkuk. Extracting a mobility model from real user traces. In *INFOCOM*, volume 6, pages 1–13, 2006. ISBN 1424402212. DOI: 10.1109/INFOCOM.2006.173.

Rob Kitchin. Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1):2053951714528481, 2014.

Niko Kiukkonen, Jan Blom, Olivier Dousse, Daniel Gatica-Perez, and Juha Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc. ICPS, Berlin*, 2010.

Tetsuo Kobayashi, Richard M Medina, and Thomas J Cova. Visualizing diurnal population change in urban areas for emergency management. *The Professional geographer : the journal of the Association of American Geographers*, 63(1):113–30, 2011. ISSN 0033-0124. URL <http://www.ncbi.nlm.nih.gov/pubmed/21491706>.

Alfred Kobsa. User acceptance of footfall analytics with aggregated and anonymized mobile phone data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8647 LNCS, pages 168–179, 2014. ISBN 9783319097695. DOI: 10.1007/978-3-319-09770-1\_15.

Martijn B W Kobus, Piet Rietveld, and Jos N. Van Ommeren. Ownership versus on-campus use of mobile it devices by university students. *Computers and Education*, 68:29–41, 2013. ISSN 03601315. DOI: 10.1016/j.compedu.2013.04.003.

Constantine E Kontokosta. The quantified community and neighborhood labs: A framework for computational urban planning and civic technology innovation. *SSRN Electronic Journal*, o(o):1–18, 2015. ISSN 1556-5068. DOI: 10.2139/ssrn.2659896. URL <http://papers.ssrn.com/abstract=2659896>.

Constantine E Kontokosta and Nicholas Johnson. Urban phenology: Toward a real-time census of the city. 2016.

Constantine E Kontokosta and Nicholas Johnson. Urban phenology: Toward a real-time census of the city using Wi-Fi data. *Computers, Environment and Urban Systems*, 64:144–153, 2017.

Gautier Krings, Francesco Calabrese, Carlo Ratti, and Vincent D Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):L07003, 2009. ISSN 1742-5468.

Johannes Kröckel and Freimut Bodendorf. Visual customer behavior analysis at the point of sale. *International Journal On Advances in Systems and Measurements*, 5(3), 2012.

John Krumm. Inference attacks on location tracks. In *International Conference on Pervasive Computing*, pages 127–143. Springer, 2007.

John Krumm. A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009.

Tarun Kulshrestha, Divya Saxena, Rajdeep Niyogi, and Jiannong Cao. Real-time crowd monitoring using seamless indoor-outdoor localization. *IEEE Transactions on Mobile Computing*, 2019.

Mei-Po Kwan and Jiyeong Lee. Emergency response after 9/11: the potential of real-time 3d gis for quick emergency response in micro-spatial environments. *Computers, Environment and Urban Systems*, 29(2):93–113, 2005. ISSN 0198-9715.

Nina Siu-Ngan Lam. Spatial interpolation methods: a review. *The American Cartographer*, 10(2):129–150, 1983. ISSN 0094-1689.

Anthony LaMarca, Yatin Chawathe, Sunny Consolvo, Jeffrey Hightower, Ian Smith, James Scott, Timothy Sohn, James Howard, Jeff Hughes, and Fred Potter. Place lab: Device positioning using radio beacons in the wild. In *International Conference on Pervasive Computing*, pages 116–133. Springer, 2005. ISBN 3540260080.

Renaud Lambiotte, Vincent D. Blondel, Cristobald de Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008. ISSN 03784371. doi: 10.1016/j.physa.2008.05.014.

Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum. *Privacy, big data, and the public good: Frameworks for engagement*. Cambridge University Press, 2014. ISBN 1316094456.

Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. A survey of mobile phone sensing. *IEEE Communications magazine*, 48(9), 2010.

Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1, 2001.

Guy Lansley and Paul Longley. Deriving age and gender from forenames for consumer analytics. *Journal of Retailing and Consumer Services*, 30:271–278, 2016a.

Guy Lansley and Paul A Longley. The geography of twitter topics in london. *Computers, Environment and Urban Systems*, 58:85–96, 2016b. ISSN 0198-9715.

Juha K Laurila, Daniel Gatica-Perez, Imad Aad, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, Markus Miettinen, and Others. The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing*, number EPFL-CONF-192489, 2012.

Bumsoo Lee. "edge"or "edgless" cities? urban spatial structure in u.s. metropolitan areas, 1980 to 2000\*. *Journal of Regional Science*, 47(3):479–515, 2007. ISSN 00224146. DOI: 10.1111/j.1467-9787.2007.00517.x.

Wouter Lefebvre, B Degrawe, Carolien Beckx, Marlies Vanhulsel, Bruno Kochan, Tom Bellemans, Davy Janssens, Geert Wets, Stijn Janssen, and Ina De Vlieger. Presentation and evaluation of an integrated model chain to respond to traffic-and health-related policy questions. *Environmental modelling & software*, 40:160–170, 2013. ISSN 1364-8152.

Songnian Li, Suzana Dragicevic, Francesc Antón Castro, Monika Sester, Stephan Winter, Arzu Coltekin, Christopher Pettit, Bin Jiang, James Haworth, Alfred Stein, et al. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS journal of Photogrammetry and Remote Sensing*, 115:119–133, 2016.

Lin Liao. *Location-based activity recognition*. PhD thesis, University of Washington, 2006.

Cristian Licoppe, Dana Diminescu, Zbigniew Smoreda, and Cezary Ziemiłki. Using mobile phone geolocalisation for 'socio-geographical' analysis of co-ordination, urban mobilities, and social integration patterns. *Tijdschrift voor Economische en Sociale Geografie*, 99(5):584–601, 2008. ISSN 0040747X. DOI: 10.1111/j.1467-9663.2008.00493.x.

Jie Lin and Robert G. Cromley. Evaluating geo-located twitter data as a control layer for areal interpolation of population. *Applied Geography*, 58: 41–47, 2015. ISSN 01436228. DOI: 10.1016/j.apgeog.2015.01.006.

Yuming Lin and Weixin Huang. A new perspective of environmental behaviour study: a brief introduction of wi-fi indoor positioning system. *International Journal of Sustainable Society*, 10(4):282–299, 2018.

Liang Liu, Clio Andris, and Carlo Ratti. Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems*, 34(6):541–548, 2010. ISSN 0198-9715.

Alyson Lloyd and James Cheshire. Detecting address uncertainty in loyalty card data. *Applied Spatial Analysis and Policy*, Jan 2018. ISSN 1874-4621. DOI: 10.1007/s12061-018-9250-1. URL <https://doi.org/10.1007/s12061-018-9250-1>.

Amy Lobben. Classification and application of cartographic animation. *The Professional Geographer*, 55(3):318–328, 2003. ISSN 0033-0124.

Sriganesh Lokanathan and Roshanthi Lucas Gunaratne. Mobile network big data for development: Demystifying the uses and challenges. *Digiworld Economic Journal*, (97):75–94, 2015. ISSN 11578637.

Ying Long and Jean-Claude Claude Thill. Combining smart card data and household travel survey to analyze jobs-housing relationships in beijing. *Computers, Environment and Urban Systems*, 53:19–35, 2015. ISSN 01989715. DOI: [10.1016/j.compenvurbsys.2015.02.005](https://doi.org/10.1016/j.compenvurbsys.2015.02.005).

Paul Longley, James Cheshire, and Alex Singleton. *Consumer Data Research*. UCL Press, 2018.

Paul A Longley, Muhammad Adnan, and Guy Lansley. The geotemporal demographics of twitter usage. *Environment and Planning A*, 47(2):465–484, 2015. ISSN 0308-518X.

Thomas Louail, Maxime Lenormand, Oliva Cantú-García, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J. Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. *arXiv:1401.4540v1 [physics.soc-ph]*, 18:1–14, 2014. ISSN 2045-2322. DOI: [10.1038/srep05276](https://doi.org/10.1038/srep05276).

Henghui Lu, Sheng Zhang, Xingchuan Liu, and Xiaokang Lin. Vehicle tracking using particle filter in wi-fi network. In *Vehicular Technology Conference Fall (VTC 2010-Fall)*, 2010 IEEE 72nd, pages 1–5. IEEE, 2010. ISBN 1424435730.

Karlo Lugomer, Balamurugan Soundararaj, Roberto Murcio, James Cheshire, and Paul Longley. Understanding sources of measurement error in the wi-fi sensor data in the smart city. In *Proceedings of GISRUK 2017*. GIS Research UK (GISRUK), 2017.

Alan M MacEachren and Menno-Jan Kraak. Research challenges in geovisualization. *Cartography and Geographic Information Science*, 28(1):3–12, 2001. ISSN 1523-0406.

Justin Manweiler, Naveen Santhapuri, Romit Roy Choudhury, and Srihari Nelakuditi. Predicting length of stay at wifi hotspots. In *INFOCOM, 2013 Proceedings IEEE*, pages 3102–3110. IEEE, 2013. ISBN 1467359440.

Stefano Marchetti, Caterina Giusti, Monica Pratesi, Nicola Salvati, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo, Luca Pappalardo, and Lorenzo Gabrielli. Small area model-based estimators using big data sources. *Journal of Official Statistics*, 31(2):263–281, 2015. ISSN 2001-7367.

David Martin. Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers*, pages 90–97, 1989. ISSN 0020-2754.

David Martin, Samantha Cockings, and Samuel Leung. Developing a flexible framework for spatiotemporal population modeling. *Annals of the Association of American Geographers*, 105(4):754–772, 2015. DOI: [10.1080/00040894.2015.1009000](https://doi.org/10.1080/00040894.2015.1009000).

10.1080/00045608.2015.1022089. URL <https://doi.org/10.1080/00045608.2015.1022089>.

Jeremy Martin, Erik Rye, and Robert Beverly. Decomposition of mac address structure for granular device inference. In *Proceedings of the 32nd Annual Conference on Computer Security Applications*, pages 78–88. ACM, 2016.

Jeremy Martin, Travis Mayberry, Collin Donahue, Lucas Foppe, Lamont Brown, Chadwick Riggins, Erik C Rye, and Dane Brown. A study of mac address randomization in mobile devices and when it fails. *arXiv preprint arXiv:1703.02874*, 2017.

Bryan H Massam. *Location and space in social administration*. E. Arnold, 1975.

Doreen Massey. Politics and space/time. *New Left Review*, I(196):249–272, 1992. ISSN 0028-6060. DOI: 10.1049/el:19990302. URL <https://newleftreview.org/I/196/doreen-massey-politics-and-space-time>.

Célestin Matte Mathieu Cunche. On wi-fi tracking and the pitfalls of mac address randomization. In *New Internet Object Challenges: Human-Machine Interaction and Human Factors*, 2016.

Célestin Matte, Mathieu Cunche, Mathy Vanhoef, Franck Rousseau, and Mathy Vanhoef. Defeating mac address randomization through timing attacks. In *ACM WiSec 2016*, pages 15–20. ACM, 2016. ISBN 1450342701.

Owen McCormack. System and method for determining demographic information, March 9 2017. US Patent App. 15/355,652.

John F McDonald. Econometric studies of urban population density: a survey. *Journal of urban economics*, 26(3):361–385, 1989. ISSN 0094-1190.

Malcolm D McIlroy, Elliot N Pinson, and Berkley A Tague. Unix time-sharing system: Foreword. *Bell System Technical Journal*, 57(6):1899–1904, 1978.

Grant McKenzie, Krzysztof Janowicz, Song Gao, and Li Gong. How where is when? on the regional variability and resolution of geosocial temporal signatures for points of interest. *Computers, Environment and Urban Systems*, 54:336–346, 2015. ISSN 0198-9715.

Dermott John James McMeel. The dark web of urban data: Fitness data ecosystems, urban design and privacy in the modern city. *International Journal of Art, Culture and Design Technologies (IJACDT)*, 7(2):12–25, 2018.

Daniel P. McMillen. Nonparametric employment subcenter identification. *Journal of Urban Economics*, 50(3):448–473, 2001. ISSN 0094-1190. DOI: 10.1006/juec.2001.2228. URL <http://www.sciencedirect.com/science/article/pii/S0094119001922284>.

Daniel P. McMillen. Employment densities, spatial autocorrelation, and subcenters in large metropolitan areas. *Journal of Regional Science*, 44(2):225–243, 2004. ISSN 00224146. DOI: 10.1111/j.0022-4146.2004.00335.x.

E Mellegard, S Moritz, and M Zahoor. Origin/destination-estimation using cellular network data. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 891–896, 2011. DOI: 10.1109/ICDMW.2011.132.

Jeremy Mennis. Generating surface models of population using dasymetric mapping. *The Professional Geographer*, 55(1):31–42, 2003. ISSN 0033-0124.

Jeremy Mennis and Torrin Hultgren. Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science*, 33(3):179–194, 2006. ISSN 1523-0406.

mikeazo and poncho. Formula for the number of expected collisions. URL <https://bit.ly/2YS6zYl>.

Harvey J. Miller. The data avalanche is here. shouldn't we be digging? *Journal of Regional Science*, 50(1):181–201, 2010. ISSN 00224146. DOI: [10.1111/j.1467-9787.2009.00641.x](https://doi.org/10.1111/j.1467-9787.2009.00641.x).

Prashanth Mohan, Venkata N Padmanabhan, and Ramachandran Ramjee. Nericell: rich monitoring of road and traffic conditions using mobile smartphones. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, pages 323–336. ACM, 2008. ISBN 1595939903.

Anna Monreale, Fabio Pinelli, Roberto Trasarti, and Fosca Giannotti. Wherenext: a location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 637–646. ACM, 2009. ISBN 1605584959.

David Moore, John Leonard, Daniela Rus, and Seth Teller. Robust distributed network localization with noisy range measurements. In *Proceedings of the 2nd international conference on Embedded networked sensor systems*, pages 50–61. ACM, 2004. ISBN 1581138792.

Julie Bauer Morrison, Barbara Tversky, and Mireille Betrancourt. Animation: Does it facilitate learning. In *AAAI spring symposium on smart graphics*, pages 53–59, 2000.

Yaser Mowafi, Ahmad Zmily, Dhiah el Diehn Abou-Tair, and Dirar Abu-Saymeh. Tracking human mobility at mass gathering events using wisp. *Future Generation Communication Technology (FGCT), 2013 Second International Conference on*, pages 157–162, 2013. DOI: [10.1109/FGCT.2013.6767212](https://doi.org/10.1109/FGCT.2013.6767212). URL [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=6767212](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6767212).

A. B. M. Musa and Jakob Eriksson. Tracking unmodified smartphones using wi-fi monitors. *Proceedings of the 10th ACM Conference on Embedded Networked Sensor Systems (SenSys '12)*, pages 281–294, 2012. DOI: [10.1145/2426656.2426685](https://doi.org/10.1145/2426656.2426685). URL <http://dl.acm.org/citation.cfm?doid=2426656.2426685>.

ABM Musa and Jakob Eriksson. Wiflow: real time travel time estimation using Wi-Fi monitors. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, pages 429–430. ACM, 2011.

Nicholas N Nagle, Barbara P Buttenfield, Stefan Leyk, and Seth Spielman. Dasymetric modeling and uncertainty. *Annals of the Association of American Geographers*, 104(1):80–95, 2014. ISSN 0004-5608.

- Taewoo Nam and Theresa A Pardo. Conceptualizing smart city with dimensions of technology, people, and institutions. In *Proceedings of the 12th annual international digital government research conference: digital government innovation in challenging times*, pages 282–291. ACM, 2011.
- F Neuhaus. Urbandiary-a tracking project. 2009. ISSN 1467-1298.
- Yibin Ng, Yingchi Pei, and Yunye Jin. Footfall count estimation techniques using mobile data. In *2017 18th IEEE International Conference on Mobile Data Management (MDM)*, pages 307–314. IEEE, 2017.
- Ivan Nikitin, Vitaly Romanov, and Giancarlo Succi. Developing wlan-based intelligent positioning system for presence detection with limited sensors. In *Guide to Ambient Intelligence in the IoT Environment*, pages 95–131. Springer, 2019.
- Eamonn O’Neill, Vassilis Kostakos, Tim Kindberg, Alan Penn, Danaë Stanton Fraser, Tim Jones, and Others. Instrumenting the city: Developing methods for observing and understanding the digital cityscape. In *International Conference on Ubiquitous Computing*, pages 315–332. Springer, 2006.
- J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007a. ISSN 0027-8424.
- Jukka-Pekka Onnela, Jari Saramäki, Jörkki Hyvönen, Gábor Szabó, M Argollo De Menezes, Kimmo Kaski, Albert-László Barabási, and János Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9(6):179, 2007b. ISSN 1367-2630.
- S. Ordonez and A. Erath. Estimating dynamic workplace capacities using public transport smart card data and a household travel survey. *Proceedings of the 17th International Conference of Hong Kong Society for Transportation Studies, HKSTS 2012: Transportation and Logistics Management*, pages 505–512, 2012. ISSN 0361-1981. DOI: 10.3141/2344-03. URL <http://www.scopus.com/inward/record.url?eid=2-s2.0-84896879480&partnerID=tZ0tx3y1>.
- Jeffrey Pang, Ben Greenstein, Ramakrishna Gummadi, Srinivasan Seshan, and David Wetherall. 802.11 user fingerprinting. In *Proceedings of the 13th annual ACM international conference on Mobile computing and networking*, pages 99–110. ACM, 2007a. ISBN 1595936815.
- Jeffrey Pang, Ben Greenstein, Srinivasan Seshan, and David Wetherall. Tryst: The case for confidential service discovery. In *HotNets*, volume 2, page 1, 2007b.
- Luca Pappalardo, Salvatore Rinzivillo, Zehui Qu, Dino Pedreschi, and Fosca Giannotti. Understanding the patterns of car travel. *The European Physical Journal Special Topics*, 215(1):61–73, 2013. ISSN 1951-6355.
- Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. *Nature Communications*, 6:8166, 2015.

ISSN 2041-1723. DOI: 10.1038/ncomms9166. URL <http://www.nature.com/ncomms/2015/150908/ncomms9166/full/ncomms9166.html>{%}5Cnhttp://www.nature.com/doifinder/10.1038/ncomms9166.

R Parrott and F.P Stutz. Urban gis applications. In P. A. Longley, M. F. Goodchild, D.J Maguire, and D.W. Rhind, editors, *Geographic information systems - Principles, techniques, applications and management*, pages 247–60. Wiley, New York, 1999.

Tao Pei, Stanislav Sobolevsky, Carlo Ratti, Shih-Lung Shaw, Ting Li, and Chenghu Zhou. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, 28(9):1988–2007, 2014. ISSN 1365-8816. DOI: 10.1080/13658816.2014.913794. URL <http://www.tandfonline.com/doi/abs/10.1080/13658816.2014.913794>.

Santi Phithakkitnukoon and Carlo Ratti. Inferring asymmetry of inhabitant flow using call detail records. *Journal of Advances in Information Technology*, 2(4):239–249, 2011. ISSN 1798-2340.

Santi Phithakkitnukoon, Teerayut Horanont, Giusy Di Lorenzo, Ryosuke Shibasaki, and Carlo Ratti. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *International Workshop on Human Behavior Understanding*, pages 14–25. Springer, 2010.

Fabio Pinelli, Giusy Di Lorenzo, and Francesco Calabrese. Comparing urban sensing applications using event and network-driven mobile phone location data. In *Mobile Data Management (MDM), 2015 16th IEEE International Conference on*, volume 1, pages 219–226. IEEE, 2015.

Juval Portugali, Han Meyer, Egbert Stolk, and Ekim Tan. *Complexity theories of cities have come of age: an overview with implications to urban planning and design*. Springer Science & Business Media, 2012. ISBN 3642245447.

R Pulselli, P Ramono, Carlo Ratti, and E Tiezzi. Computing urban mobile landscapes through monitoring population density based on cellphone chatting. *Int. J. of Design and Nature and Ecodynamics*, 3(2):121–134, 2008.

Rami Puzis, Yaniv Altshuler, Yuval Elovici, Shlomo Bekhor, Yoram Shiftan, and Alex Pentland. Augmented betweenness centrality for environmentally aware traffic monitoring in transportation networks. *Journal of Intelligent Transportation Systems*, 17(1):91–105, 2013. ISSN 1547-2450.

Weijun Qin, Jiadi Zhang, Bo Li, and Limin Sun. Discovering human presence activities with smartphones using nonintrusive wi-fi sniffer sensors: The big data prospective. *International Journal of Distributed Sensor Networks*, 2013, 2013. ISSN 15501477. DOI: 10.1155/2013/927940.

John N K Rao and Isabel Molina. *Small area estimation*. John Wiley & Sons, 2015. ISBN 1118735722.

Carlo Ratti, Dennis Frenchman, Riccardo Maria Pulselli, and Sarah Williams. Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748, 2006. ISSN 02658135. DOI: 10.1068/b32047.

- Jonathan Reades, Francesco Calabrese, Andres Sevtsuk, and Carlo Ratti. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, 6(3):30–38, 2007. ISSN 15361268. DOI: 10.1109/MPRV.2007.53.
- Jonathan Reades, Francesco Calabrese, and Carlo Ratti. Eigenplaces: Analysing cities using the space - time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5):824–836, 2009. ISSN 02658135. DOI: 10.1068/b34133t.
- Michael Reibel and Michael E Bufalino. Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A*, 37(1):127–139, 2005. ISSN 0308-518X.
- Jun Rekimoto, Takashi Miyaki, and Takaaki Ishizawa. Lifetag: Wifi-based continuous location logging for life pattern analysis. In *LoCA*, volume 2007, pages 35–49, 2007.
- Ian Rose and Matt Welsh. Mapping the urban wireless landscape with argos. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, pages 323–336. ACM, 2010. ISBN 1450303447.
- Luca Rossi, James Walker, and Mirco Musolesi. Spatio-temporal techniques for user identification by means of gps mobility data. *EPJ Data Science*, 4(1):11, 2015.
- Priya Roy and Chandreyee Chowdhury. Indoor localization for smart-handhelds with stable set of wireless access points. In *2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT)*, pages 1–4. IEEE, 2018a.
- Priya Roy and Chandreyee Chowdhury. Smartphone based indoor localization using stable access points. In *Proceedings of the Workshop Program of the 19th International Conference on Distributed Computing and Networking*, page 17. ACM, 2018b.
- Günther Sagl, Bernd Resch, Bartosz Hawelka, and Euro Beinat. From social sensor data to collective human behaviour patterns: Analysing and visualising spatio-temporal dynamics in urban environments. In *Proceedings of the GI-Forum*, pages 54–63, 2012.
- Piotr Sapiezynski, Arkadiusz Stopczynski, Radu Gatej, and Sune Lehmann. Tracking human mobility using wifi signals. *PloS one*, 10(7):e0130824, 2015. ISSN 1932-6203. DOI: 10.1371/journal.pone.0130824. URL <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130824>.
- T Scott Saponas, Jonathan Lester, Carl Hartung, Sameer Agarwal, and Tadayoshi Kohno. Devices that tell on you: Privacy trends in consumer ubiquitous computing. In *Usenix Security*, volume 3, page 3, 2007.
- Saskia Sassen. *The global city: New York, London, Tokyo*, volume 15. 2001. ISBN 9780691070636. DOI: 10.2307/2152688. URL <http://books.google.com/books?id=PTAiHwK2BYIC>.

Marco Luca Sbodio, Francesco Calabrese, Michele Berlingero, Rahul Nair, and Fabio Pinelli. All aboard: visual exploration of cellphone mobility data to optimise public transport. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 335–340. ACM, 2014. ISBN 1450321844.

Lorenz Schauer, Martin Werner, and Philipp Marcus. Estimating crowd densities and pedestrian flows using wi-fi and bluetooth. *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 171–177, 2014. DOI: 10.4108/icst.mobiquitous.2014.257870. URL <http://eudl.eu/doi/10.4108/icst.mobiquitous.2014.257870>.

Johannes Schlaich, Thomas Otterstätter, and Markus Friedrich. Generating trajectories from mobile phone data. In *Proceedings of the 89th annual meeting compendium of papers, transportation research board of the national academies*, 2010.

Robert C Schmitt. Estimating daytime populations. *Journal of the American Institute of Planners*, 22(2):83–85, 1956. ISSN 0002-8991.

Andres Sevtsuk and Carlo Ratti. Does urban mobility have a daily routine? learning from the aggregate data of mobile networks. *Journal of Urban Technology*, 17(1):41–60, 2010. ISSN 1063-0732. DOI: 10.1080/10630731003597322. URL <http://www.tandfonline.com/doi/abs/10.1080/10630731003597322>.

Kyosuke Shibata and Hiroshi Yamamoto. People crowd density estimation system using deep learning for radio wave sensing of cellular communication. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, pages 143–148. IEEE, 2019.

Masamichi Shimosaka, Takuhiro Kaneko, Kentaro Nishi, Masamichi Shimosaka, Kentaro Nishi, Junichi Sato, and Hirokatsu Kataoka. Extracting land-use patterns using location data from smartphones. *The 1st International Conference on IoT in Urban Space*, (Iv):1694–1700, 2014. DOI: 10.4108/icst.urb-iot.2014.257220.

Dongyoun Shin, Sofia Georagakopoulou, Daniel Zünd, and Gerhard Schmitt. Crowdsourcing urban sensing: Mobile phone for urban data collection. In *2nd International Hybrid City Conference*, pages 25–28, 2013. ISBN 978-960-99791-1-5.

Seung Hyuck Shin, Chan Gook Park, and Sangon Choi. New map-matching algorithm using virtual track for pedestrian dead reckoning. *ETRI journal*, 32(6):891–900, 2010.

Lei Shu, Yuanfang Chen, Zhiqiang Huo, Neil Bergmann, and Lei Wang. When mobile crowd sensing meets traditional industry. *IEEE Access*, 5:15300–15307, 2017.

Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012. ISSN 0028-0836.

Alex David Singleton and Paul Longley. The internal structure of greater london: a comparison of national and regional geodemographic models. *Geo: Geography and Environment*, 2(1):69–87, 2015. ISSN 2054-4049.

- Alan Smith, David Martin, and Samantha Cockings. Spatio-temporal population modelling for enhanced assessment of urban exposure to flood risk. *Applied Spatial Analysis and Policy*, 9(2):145–163, 2016. ISSN 1874-463X.
- Edward Soja. Postmodern geographies, 1989. URL <http://books.google.com/books?id=sNcRAQAAIAAJ>.
- Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010a. ISSN 1745-2473. DOI: [10.1038/nphys1760](https://doi.org/10.1038/nphys1760).
- Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010b. ISSN 0036-8075, 1095-9203. DOI: [10.1126/science.1177170](https://doi.org/10.1126/science.1177170).
- Balamurugan Soundararaj. Clicker - an Android app for counting pedestrian footfalls with accuracy. <https://play.google.com/store/apps/details?id=com.bala.manualcount>, 2018.
- MS Sruthi. IoT based real time people counting system for smart buildings. *International Journal of Emerging Technology and Innovative Engineering*, 5(2), 2019.
- John Steenbruggen, Maria Teresa Borzacchiello, Peter Nijkamp, and Henk Scholten. Mobile phone data from gsm networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. *GeoJournal*, 78(2):223–243, 2013. ISSN 0343-2521.
- John Steenbruggen, Emmanouil Tranos, and Peter Nijkamp. Data from mobile phone operators: A tool for smarter cities? *Telecommunications Policy*, 39(3-4):335–346, 2015. ISSN 03085961. DOI: [10.1016/j.telpol.2014.04.001](https://doi.org/10.1016/j.telpol.2014.04.001).
- Shan Suthaharan. Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review*, 41(4):70–73, 2014.
- Paul Sutton. Modeling population density with night-time satellite imagery and gis. *Computers, Environment and Urban Systems*, 21(3):227–244, 1997. ISSN 0198-9715.
- Paul Sutton, Dar Roberts, C Elvidge, and Kimberly Baugh. Census from heaven: An estimate of the global human population using night-time satellite imagery. *International Journal of Remote Sensing*, 22(16):3061–3076, 2001. ISSN 0143-1161.
- Karen P Tang, Pedram Keyani, James Fogarty, and Jason I Hong. Putting people in their place: an anonymous and privacy-sensitive approach to collecting sensed data in location-based applications. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 93–102. ACM, 2006.
- Ole Tange. *GNU Parallel 2018*. Ole Tange, March 2018. ISBN 9781387509881. DOI: [10.5281/zenodo.1146014](https://doi.org/10.5281/zenodo.1146014). URL <https://doi.org/10.5281/zenodo.1146014>.

Anna F Tapp. Areal interpolation and dasymetric mapping methods using local ancillary data sources. *Cartography and Geographic Information Science*, 37(3):215–228, 2010. ISSN 1523-0406.

Nicholas Taylor, Richelle Mayshak, Ashlee Curtis, Kerri Coomber, Tanya Chikritzhs, and Peter Miller. Investigating and validating methods of monitoring foot-traffic in night-time entertainment precincts in australia. *International Journal of Drug Policy*, 66:23–29, 2019.

Deborah S K Thomas. Data, data everywhere, but can we really use them. *American hazardscapes: The regionalization of hazards and disasters*, pages 61–76, 2001.

James J Thomas. *Illuminating the path:[the research and development agenda for visual analytics]*. IEEE Computer Society, 2005. ISBN 0769523234.

Waldo R Tobler. Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74(367):519–530, 1979. ISSN 0162-1459.

Jameson L. Toole, Michael Ulm, Marta C. González, and Dietmar Bauer. Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing - UrbComp '12*, page 1, 2012. ISBN 9781450315425. DOI: 10.1145/2346496.2346498. URL <http://dl.acm.org/citation.cfm?doid=2346496.2346498>.

Paul M. Torrens. Wi-fi geographies. *Annals of the Association of American Geographers*, 98(1):59–84, 2008. DOI: 10.1080/00045600701734133.

Anthony M. Townsend. Life in the real-time city: Mobile telephones and urban metabolism. *Journal of Urban Technology*, 7(2):85–104, 2000. ISSN 1063-0732. DOI: 10.1080/713684114.

Emmanouil Tranos. *The geography of the internet: Cities, regions and internet infrastructure in Europe*. Edward Elgar Publishing, 2013.

Emmanouil Tranos and Peter Nijkamp. Urban and regional analysis and the digital revolution: challenges and opportunities. *Hub cities in the knowledge economy: Ashgate, Forthcoming*, 2012.

Emmanouil Tranos and Peter Nijkamp. Mobile phone usage in complex urban systems: a space- time, aggregated human activity study. *Journal of Geographical Systems*, 17(2):157–185, 2015. ISSN 14355949. DOI: 10.1007/s10109-015-0211-9.

Emmanouil Tranos, John Steenbruggen, and Peter Nijkamp. 8 mobile phone operators, their (big) data and urban analysis. *A Research Agenda for Regeneration Economies: Reading City-Regions*, page 140, 2018.

Toivo Vajakas, Tanel Kiis, Amnir Hadachi, and Eero Vainikko. Mobility episode discovery in the mobile networks based on enhanced switching kalman filter. In *2018 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pages 1–7. IEEE, 2018.

Mathy Vanhoef, Célestin Matte, Mathieu Cunche, Leonardo S Cardoso, Frank Piessens, and Piessens Frank. Why mac address randomization is not enough: An analysis of wi-fi network discovery mechanisms. In *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*, pages 413–424. ACM, 2016. ISBN 1450342337. DOI: [10.1145/2897845.2897883](https://doi.org/10.1145/2897845.2897883).

Leah K VanWey, Ronald R Rindfuss, Myron P Gutmann, Barbara Entwistle, and Deborah L Balk. Confidentiality and spatially explicit data: Concerns and challenges. *Proceedings of the National Academy of Sciences*, 102(43):15337–15342, 2005. ISSN 0027-8424.

Tien Dang Vo-Huu, Triet Dang Vo-Huu, and Guevara Noubir. Fingerprinting Wi-Fi devices using software defined radios. In *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, pages 3–14. ACM, 2016.

Long Vu, Quang Do, and Klara Nahrstedt. Jyotish: A novel framework for constructing predictive model of people movement from joint wifi/bluetooth trace. In *Pervasive Computing and Communications (PerCom), 2011 IEEE International Conference on*, pages 54–62. IEEE, 2011.

Stephen R. Walli. The posix family of standards. *StandardView*, 3(1):11–17, March 1995. ISSN 1067-9936. DOI: [10.1145/210308.210315](https://doi.acm.org/10.1145/210308.210315). URL <http://doi.acm.org/10.1145/210308.210315>.

Neng Wan and Ge Lin. Life-space characterization from cellular telephone collected gps data. *Computers, Environment and Urban Systems*, 39:63–70, 2013. ISSN 0198-9715.

D. Wang, D. Pedreschi, C. Song, Fosca Giannotti, and Albert-László Barabási. Human mobility, social ties, and link prediction. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM., pages 100–1108, 2011. ISBN 1450308139. DOI: [10.1145/2020408.2020581](https://doi.org/10.1145/2020408.2020581).

Pu Wang, Timothy Hunter, Alexandre M Bayen, Katja Schechtner, and Marta C González. Understanding road usage patterns in urban areas. *Scientific reports*, 2, 2012.

W. Wang, A.X. Liu, and M. Shahzad. Gait recognition using wifi signals. In *UbiComp 2016 - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016. ISBN 9781450344616. DOI: [10.1145/2971648.2971670](https://doi.org/10.1145/2971648.2971670).

Yan Wang, Jie Yang, Hongbo Liu, and Yingying Chen. Measuring human queues using wifi signals. In *Proceedings of the 19th annual international conference on Mobile computing & networking*, pages 235–237, 2013. ISBN 9781450319997. DOI: [10.1145/2500423.2504584](https://doi.org/10.1145/2500423.2504584). URL <http://dl.acm.org/citation.cfm?doid=2500423.2504584{\%}5Cnhttp://dl.acm.org/citation.cfm?id=2504584>.

Amy Wesolowski, Nathan Eagle, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society Interface*, 10(81):20120986, 2013. ISSN 1742-5689.

Martin Wirz, Tobias Franke, Daniel Roggen, Eve Mitleton-Kelly, Paul Lukowicz, and Gerhard Tröster. Inferring crowd conditions from pedestrians' location traces for real-time crowd monitoring during city-scale mass gatherings. In *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2012 IEEE 21st International Workshop on*, pages 367–372. IEEE, 2012. ISBN 1467318884.

Luc Johannes Josephus Wismans, LO de Vries, and EC van Berkum. Comparison of interpolation techniques for state estimation on urban networks. In *NECTAR conference 2017*, 2017.

Wei Xi, Yuan He, Yunhao Liu, Jizhong Zhao, Lufeng Mo, Zheng Yang, Jiliang Wang, and Xiangyang Li. Locating sensors in the wild: pursuit of ranging quality. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, pages 295–308. ACM, 2010. ISBN 1450303447.

Zhuliang Xu, Kumbesan Sandrasegaran, Xiaoying Kong, Xinning Zhu, Jingbin Zhao, Bin Hu, and Cheng-Chung Lin. Pedestrain monitoring system using Wi-Fi technology and rssi based localization. *International Journal of Wireless & Mobile Networks*, 5(4):17, 2013.

Yingxiang Yang, David Gerstle, Peter Widhalm, Dietmar Bauer, and Marta Gonzalez. Potential of low-frequency automated vehicle location data for monitoring and control of bus performance. *Transportation Research Record: Journal of the Transportation Research Board*, (2351):54–64, 2013. ISSN 0361-1981.

Yew Yuan, Richard M Smith, and W Fredrick Limp. Remodeling census population with spatial information from landsat tm imagery. *Computers, Environment and Urban Systems*, 21(3):245–258, 1997. ISSN 0198-9715.

Yihong Yuan and Martin Raubal. Analyzing the distribution of human activity space from mobile phone usage: an individual and urban-oriented study. *International Journal of Geographical Information Science*, 30(8):1594–1621, 2016. ISSN 1365-8816. DOI: 10.1080/13658816.2016.1143555. URL <http://www.tandfonline.com/doi/full/10.1080/13658816.2016.1143555>.

Sakir Yucel. Measuring benefits, drawbacks and risks of smart community wireless platforms. In *Proceedings of the International Conference on Modeling, Simulation and Visualization Methods (MSV)*, pages 107–113. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2017.

Yang Yue, Han-dong Wang, Bo Hu, Qing-quan Li, Yu-guang Li, and Anthony G O Yeh. Exploratory calibration of a spatial interaction model using taxi gps trajectories. *Computers, Environment and Urban Systems*, 36(2):140–153, 2012. ISSN 0198-9715.

Robbert Zandvliet and Martin Dijst. Short-term dynamics in the use of places: A space-time typology of visitor populations in the netherlands. *Urban Studies*, 43(7):1159–1176, 2006.

Vasileios Zarimpas, Bahram Honary, and Mike Darnell. Indoor 802.1 x based location determination and real-time tracking. In *Wireless, Mobile and Multimedia Networks, 2006 IET International Conference on*, pages 1–4. IET, 2006. ISBN 0863416446.

Zengrzengr, Andy, and Cabral. Calculate distance from rssi. URL <https://bit.ly/20hskf9>.

Desheng Zhang, Jun Huang, Ye Li, Fan Zhang, Chengzhong Xu, and Tian He. Exploring human mobility with multi-source data at extremely large metropolitan scales. *Proceedings of the 20th annual international conference on Mobile computing and networking - MobiCom '14*, pages 201–212, 2014. DOI: 10.1145/2639108.2639116. URL <http://dl.acm.org/citation.cfm?id=2639108.2639116>.

Kaisa Zhang, Gang Chuai, Weidong Gao, Xuewen Liu, Saidiwaerdi Maimaiti, and Zhiwei Si. A new method for traffic forecasting in urban wireless communication network. *EURASIP Journal on Wireless Communications and Networking*, 2019(1):66, 2019.

Yanting Zhang, Shuaiyu Jin, Yuanyuan Qiao, Kewu Sun, Hao Zhang, and Jie Yang. Exploring urban spatial hotspots' properties using inter-connected user-location networks. In *2018 IEEE/CIC International Conference on Communications in China (ICCC)*, pages 84–89. IEEE, 2018.

Huijing Zhao and Ryosuke Shibasaki. A novel system for tracking pedestrians using multiple single-row laser-range scanners. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 35(2):283–291, 2005. ISSN 1083-4427.

Chen Zhong, Michael Batty, Ed Manley, Jiaqiu Wang, Zijia Wang, Feng Chen, and Gerhard Schmitt. Variability in regularity: Mining temporal mobility patterns in london, singapore and beijing using smart-card data. *PLoS ONE*, 11(2), 2016. ISSN 19326203. DOI: 10.1371/journal.pone.0149222.

Enwei Zhu, Maham Khan, Philipp Kats, Shreya Santosh Bamne, and Stanislav Sobolevsky. Digital urban sensing: A multi-layered approach. *arXiv preprint arXiv:1809.01280*, 2018.